Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) – Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Алгоритмы сравнения схожести текстов на основе нейросетевых алгоритмов

УДК 004.032.26:81

Студент

Группа	ФИО	Подпись	Дата
8ВМ13	Цыденов Саян Баирович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

Консультант

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент	Кривошеев Н.А.	аспирант		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Былкова Т.В.	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Федорчук Ю.М.	д.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

Томск – 2023г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП

по направлению 09.04.01 Информатика и вычислительная техника

Код компетенции	**Наименование компетенции**
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований

ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен разрабатывать компоненты программно-аппаратных комплексов обработки информации и автоматизированного проектирования
ОПК(У)-7	Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов
ПК(У)-2	Способен проектировать сложные пользовательские интерфейсы
ПК (У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК (У)-4	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ
ПК(У)-5	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) – Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Спицын В.Г.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ВМ13	Цыденов Саян Баирович

Тема работы:

Алгоритмы сравнения схожести текстов на основе нейросетевых алгоритмов	
Утверждена приказом директора (дата, номер)	№ 40-57/с от 09.02.2023

Срок сдачи студентом выполненной работы:	12.06.2023

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Объектом проектирования и разработки является нейросетевой алгоритм для семантического сравнения текстов.
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Анализ и обзор существующих методов семантической сегментации облаков точек; 2. Проектирование и разработка нейросетевого алгоритма; 3. Тестирование нейросетевого алгоритма; 4. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения;

	5. Работа над разделом по социальной ответственности; 6. Работа над разделом на английском языке.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Изображения примеров из набора данных; 2. Схемы структур нейронных сетей; 3. Изображения диаграмм рассеяния. 4. Таблицы для сравнения моделей; 5. Предсказанные метки классов для тестовых выборок.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.э.н., Былкова Т.В.
Социальная ответственность	Профессор ООД ШБИП, д.т.н., Федорчук Ю.М.
Английский язык	Доцент ОИЯ ШБИП, к.п.н., Сидоренко Т.В.

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Раздел 2 Text similarity comparison algorithms based on neural network algorithms

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	01.03.2023

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ13	Цыденов Саян Баирович		

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) – Отделение информационных технологий
Период выполнения (осенний / весенний семестр 2022 /2023 учебного года)

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	12.06.2023

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
12.06.2022	Анализ и обзор существующих методов семантического сравнения текстов	20
12.06.2022	Проектирование и разработка нейросетевого алгоритма	30
12.06.2022	Тестирование нейросетевого алгоритма	20
12.06.2022	Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения	10
12.06.2022	Работа над разделом по социальной ответственности	10
12.06.2022	Работа над разделом на английском языке	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ВМ13	Цыденов Саян Баирович

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	магистратура	Направление/специальность	09.04.01 «Информатика и вычислительная техника»

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	Стоимость ресурсов определялась по средней рыночной стоимости, и в соответствии с окладами сотрудников организации.
2. *Нормы и нормативы расходования ресурсов*	Районный коэффициент 30%, коэффициент дополнительной заработной платы 12%;
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	Коэффициент отчислений на уплату во внебюджетные фонды 30%.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	Провести предпроектный анализ
2. *Разработка устава научно-технического проекта*	Представить Устав научного проекта магистерской работы
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	Разработать план управления НТИ
4. *Определение ресурсной, финансовой, экономической эффективности*	Рассчитать сравнительную эффективность исследования
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *Карта сегментирования рынка* 2. *Оценка конкурентоспособности технических решений* 3. *Матрица SWOT* 4. *Оценка степени готовности проекта к коммерциализации* 5. *Заинтересованные стороны проекта* 6. *Цели и результат проекта* 7. *Организационная структура проекта* 8. *Календарный план-график проведения НИОКР по теме* 9. *Баланс рабочего времени участников разработки* 10. *Расчет основной заработной платы* 11. *Бюджет затрат НТИ* 12. *Сравнение характеристик вариантов исполнения проекта* 13. *Сравнительные показатели эффективности разработки*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Былкова Т.В.	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ13	Цыденов Саян Баирович		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа		**ФИО**	
8ВМ13		Цыденов Саян Баирович	
Школа	ИШИТР	**Отделение(НОЦ)**	ОИТ
Уровень образования	магистратура	**Направление/ специальность**	*09.04.01«Информатикаи вычислительнаятехника»*

Тема ВКР:

Алгоритмы сравнения схожести текстов на основе нейросетевых алгоритмов	
Исходныеданныекразделу«Социальнаяответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/приэксплуатации	*Объект исследования Программный модуль семантического сравнения текстов* *Область применения: Поисковые системы* *Рабочая зона:* офис *Размеры помещения: Ширина комнаты составляетb=3м, длинаа=6м, высота H = 3 м. Площадь помещения будет составлять S=ab=18*м², *объем V=abh=54* м³*;* *Количество и наименование оборудования рабочей зоны: персональный компьютер, устройства ввода и вывода информации, ЖК монитор.* *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне: присутствует окно, через которое может производится вентиляция помещения, принудительная вентиляция отсутствует; в зимнее время помещение отапливается; впомещении используется комбинированное освещение.*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Производственная безопасность 1.1. Анализ выявленных вредных факторов • Природавоздействия • Действие на организмчеловека • Нормы воздействия и нормативные документы (для вредных факторов) • СЗ коллективные и индивидуальные 1.2. Анализвыявленныхопасныхфакторов: • Термическиеисточникиопасности • Электроопасность • Пожароопасности	1. Вредныефакторы: 1.1. Недостаточнаяосвещенность; 1.2. Нарушения микроклимата, оптимальные и допустимые параметры; 1.3. Шум, ПДУ, СКЗ, СИЗ; 1.4. Повышенный уровень электромагнитного излучения, ПДУ, СКЗ, СИЗ; 2.1. Электроопасность; класс электроопасности помещения, безопасные номиналы I, U, $R_{заземления}$, СКЗ, СИЗ; Приведен расчет освещения рабочего места; 2.2. Пожароопасность, категория пожароопасности помещения, марки огнетушителей, их назначение и ограничение применения; Приведена схема эвакуации.

2. Экологическаябезопасность: • Выбросы в окружающуюсреду • Решенияпообеспечениюэкологическойбезопасности	Наличие промышленных отходов (бумага-черновики, вторцвет и чермет, пластмасса, перегоревшие люминесцентные лампы, оргтехника) и способы их утилизации;
3. Безопасность в чрезвычайных ситуациях: 1.перечень возможных ЧС при разработке и эксплуатации проектируемого решения; 2.разработка превентивных мер по предупреждению ЧС; 3.разработка действий в результате возникшей ЧС и мер по ликвидации её последствий.	Рассмотрены 2 ситуации ЧС: 1) природная – сильные морозы зимой, (аварии на электро-, тепло-коммуникациях, водоканале, транспорте); 2) техногенная – несанкционированное проникновение посторонних на рабочее место (возможны проявления вандализма, диверсии, промышленного шпионажа), представлены мероприятия по обеспечению устойчивой работы производства в том и другом случае.
4. Переченьнормативно-техническойдокументации.	– ГОСТы, СанПиНы, СНиПы

Дата выдачи задания для раздела по линейному графику	22.05.2023 г.

Задание выдал консультант:

Должность	ФИО	Ученаястепеньзвание	Подпись	Дата
Профессор ООД ШБИП	Федорчук Ю.М..	д.т.н.		22.05.2023 г.

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ13	Цыденов Саян Баирович		22.05.2023 г.

Реферат

Выпускная квалификационная работа содержит пояснительную записку на 94 страницах, 12 рисунков, 22 таблицы, 49 источников литературы, 1 приложение.

Ключевые слова: машинное обучение, семантическое сравнение текстов, алгоритмы обработки естественного языка, архитектура BERT, трансформеры в нейронных сетях.

Целью данной работы является реализация алгоритма машинного обучения для сравнения текстов по их смыслу. Для достижения данной цели необходимо: проанализировать существующие методы семантического сравнения текстов, выбрать подходящую модель для реализации, выбрать набор данных, реализовать модель, произвести обучение и тестирование.

В рамках исследования проводился анализ существующих нейросетевых моделей и методов для семантического сравнения текстов, на основании которого была проведена выборка модели для дальнейшей реализации.

В результате был предложен алгоритм семантического сравнения текстов, основанный на архитектуре BERT и методах трансформеров. Была разработана модель, основанная на BERT, включающая в себя использование косинусного и евклидового расстояний, а также обученной нейронной сети по векторам. Проведение численных экспериментов осуществлялось на выбранных наборах данных. В результате тестирования были получены обнадеживающие результаты по качеству семантического сопоставления текстов.

Было проведено сравнение с другими существующими методами семантического сравнения текстов, результат которого показал, что реализованная модель демонстрирует высокие результаты в задачах семантического сравнения, подтверждая эффективность выбранного подхода.

Оглавление

Введение

С развитием информационных технологий и увеличением объемов доступных данных, возникла необходимость в эффективных методах анализа и обработки текстовой информации. Одной из ключевых задач в этой области является семантическое сравнение текстов, то есть определение степени схожести двух текстов по их смыслу. Эта задача имеет широкий спектр применений, включая поиск информации, системы рекомендаций, анализ тональности.

Целью данной работы является разработка и реализация алгоритма машинного обучения для решения задачи семантического сравнения текстов. В основе предлагаемого подхода лежат методы нейронных сетей, в частности, архитектура трансформеров и модель BERT (Bidirectional Encoder Representations from Transformers), которые демонстрируют высокую эффективность в задачах обработки естественного языка.

В работе рассмотрено несколько методов семантического сравнения текстов, включая сравнение по косинусным и евклидовым расстояниям, а также с использованием обученной нейронной сети по векторам. Каждый из этих методов имеет свои преимущества и недостатки, и выбор конкретного метода зависит от специфики задачи и доступных данных.

Важной частью работы является проведение экспериментов для оценки эффективности предложенного алгоритма. Были проведены эксперименты на различных наборах данных. Результаты были сравнены с результатами применения других существующих методов.

1. Описание проблемы и актуальности темы

Семантическое сравнение текстов является одной из ключевых задач в области обработки естественного языка (NLP). Она заключается в определении степени схожести двух текстов по их смыслу, что представляет собой сложную и многогранную проблему. С одной стороны, тексты могут быть схожими по структуре и словарному запасу, но отличаться по смыслу. С другой стороны, тексты могут быть написаны разными словами и стилями, но передавать одну и ту же идею. Таким образом, задача семантического сравнения требует учета как лексических, так и семантических аспектов текста.

Актуальность темы обусловлена растущими потребностями в автоматизации обработки текстовой информации. Семантическое сравнение текстов играет важную роль во многих приложениях, таких как поиск информации, системы рекомендаций, анализ тональности, машинный перевод и многие другие. Например, в поисковых системах семантическое сравнение может помочь улучшить качество поиска, позволяя системе лучше понимать запросы пользователей и предлагать более релевантные результаты. В системах рекомендаций семантическое сравнение может помочь предлагать пользователям контент, который более точно соответствует их интересам.

Несмотря на значительные успехи в области NLP, семантическое сравнение текстов все еще остается сложной задачей, требующей дальнейших исследований. В частности, существующие методы могут столкнуться с проблемами при работе с текстами, содержащими сложные семантические структуры, неоднозначности или специфический доменный словарный запас. В этой работе разработан алгоритм семантического сравнения текстов, основанный на методах машинного обучения, который будет способен справиться с этими и другими вызовами.

2. Теоретический обзор

2.1. Обзор методов машинного обучения для семантического сравнения текстов

Семантическое сравнение текстов – это задача, в которой машинное обучение играет ключевую роль. Методы машинного обучения, применяемые для этой задачи, варьируются от классических подходов, таких как мешок слов (bag of words) и TF-IDF, до более сложных и современных методов, основанных на нейронных сетях.

Мешок слов и TF-IDF являются простыми, но эффективными методами для преобразования текста в векторное пространство, что позволяет использовать стандартные метрики сходства. Однако эти методы не учитывают порядок слов и семантические отношения между ними, что ограничивает их способность к семантическому сравнению текстов.

Нейронные сети представляют собой более мощный инструмент для семантического сравнения текстов. Они способны моделировать сложные зависимости и семантические структуры в тексте, что делает их особенно полезными для этой задачи. Одним из ключевых преимуществ нейронных сетей является их способность к обучению на больших объемах данных, что позволяет им извлекать более глубокие и точные семантические представления текста.

Архитектура трансформеров, предложенная в работе "Attention is All You Need" (Vaswani et al., 2017) [1], представляет собой значительный прорыв в области обработки естественного языка. Трансформеры используют механизм внимания для моделирования зависимостей между словами в тексте, что позволяет им эффективно обрабатывать длинные последовательности и учитывать контекст каждого слова. Это делает их особенно подходящими для задач семантического сравнения текстов.

BERT (Bidirectional Encoder Representations from Transformers) – это модель, основанная на архитектуре трансформеров, которая была представлена в работе «BERT: Pre-training of Deep Bidirectional Transformers

for Language Understanding» (2018) [2]. BERT использует двунаправленное обучение на представлениях трансформеров, что позволяет ему лучше понимать контекст и семантику слов.

2.2. Обзор архитектуры трансформеров

Архитектура трансформеров была впервые представлена в 2017 году и с тех пор стала основой для многих современных моделей обработки естественного языка. Основной идеей трансформеров является использование механизма внимания, который позволяет моделировать зависимости между словами в тексте без учета их позиции в последовательности.

Трансформеры состоят из двух основных частей: энкодера и декодера. Энкодер преобразует входной текст в последовательность внутренних векторных представлений, каждое из которых отражает смысл соответствующего слова в контексте всего текста. Декодер затем использует эти представления для генерации выходного текста, также учитывая контекст.

Основной компонент трансформеров — это блок внимания, который вычисляет веса взаимодействия между всеми парами слов в тексте. Это позволяет модели учитывать контекст каждого слова, независимо от его позиции в тексте. Благодаря этому трансформеры могут эффективно обрабатывать длинные последовательности и улавливать сложные семантические зависимости между словами.

Трансформеры также используют механизм позиционного кодирования для передачи информации о порядке слов в тексте. Это позволяет модели учитывать порядок слов несмотря на то, что сам механизм внимания не учитывает позицию слов.

С момента своего появления архитектура трансформеров стала основой для многих других моделей, таких как BERT, GPT-4 и T5, которые демонстрируют выдающиеся результаты в различных задачах обработки естественного языка.

2.3. Обзор архитектуры BERT и ее применения для семантического сравнения текстов

BERT (Bidirectional Encoder Representations from Transformers) — это модель обработки естественного языка [2]. BERT основан на архитектуре трансформеров и использует двунаправленное обучение для создания мощных представлений текста.

В отличие от некоторых предшествующих моделей, таких как GPT, которые обучаются предсказывать следующее слово в тексте (так называемое "обучение с учителем"), BERT обучается на двух задачах: предсказании пропущенных слов в тексте (задача "заполнения пробелов") и определении, является ли одно предложение продолжением другого. Это двунаправленное обучение позволяет BERT лучше понимать контекст и семантику слов, что делает его особенно полезным для задач семантического сравнения текстов.

BERT обучается на большом корпусе текстов и создает векторное представление для каждого слова в тексте. Эти векторные представления затем могут быть использованы для семантического сравнения текстов, например, путем вычисления косинусного или евклидового расстояния между векторами. BERT также может быть дообучен на конкретной задаче, что позволяет ему адаптироваться к специфическим требованиям задачи и улучшить его производительность.

Применение BERT для семантического сравнения текстов уже демонстрировало впечатляющие результаты в ряде задач, включая определение семантической эквивалентности предложений, классификацию текстов и ранжирование документов. Это делает BERT одним из наиболее мощных и гибких инструментов для семантического сравнения текстов на сегодняшний день.

2.3.1. Архитектура BERT

Архитектура BERT представляет собой двунаправленный предварительно обученный энкодер, способный эффективно моделировать контекстуальные зависимости в тексте. Основная идея заключается в

предварительном обучении модели на большом объеме неразмеченных текстовых данных и использовании полученных весов для дальнейшего обучения на конкретных задачах NLP.

Преимущество BERT заключается в его способности учиться контекстуальным представлениям слов, то есть понимать значение слова в зависимости от его окружения в предложении. Это достигается благодаря механизму self-attention, который позволяет модели учитывать взаимодействия между различными словами в предложении.

BERT обучается на двух задачах: предсказание следующего предложения и маскирование слов. В первой задаче модель обучается предсказывать, является ли одно предложение продолжением другого. Во второй задаче случайным образом выбираются слова в предложении и заменяются маской или случайным словом, а модель должна предсказать исходное слово. На рисунке 2.1 приведено входное представление BERT.

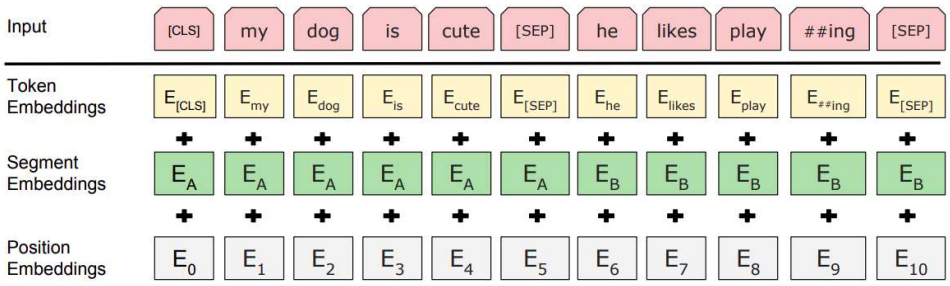

Рисунок 2.1. BERT - входное представление.

После предварительного обучения на неразмеченных данных, сеть BERT может быть дообучена на различных задачах NLP, таких как классификация текста, извлечение информации, вопросно-ответные системы и многое другое. Однако, дообучение модели требует размеченных данных для конкретной задачи.

Сеть BERT достигла высоких результатов во многих задачах NLP и стала стандартным инструментом для многих исследований и приложений. Она предоставляют мощный инструмент для работы с естественным языком и

позволяют извлекать богатые семантические представления из текстовых данных.

2.3.2. Основные элементы архитектуры BERT:

- **Внутреннее векторное представление слов**: Входной текст разбивается на отдельные слова, называемые токенами. Каждый токен представляется векторным представлением, называемым внутренним представлением слова. В BERT используется внутренние представления слов, которые могут быть предварительно обучены или обучены вместе с моделью.
- **Сегментные внутренние представления**: при обработке пар предложений, BERT использует специальные сегментные внутренние представления для разделения и отделения предложений друг от друга. Каждый токен отмечается сегментным идентификатором, указывающим принадлежность к определенному предложению.
- **Позиционные внутренние представления**: для учета позиционной информации в тексте, BERT использует позиционные внутренние представления. Они представляют относительные позиции токенов в предложении и позволяют модели учитывать порядок слов.
- **Многоуровневые трансформеры**: Основная часть архитектуры BERT состоит из нескольких слоев трансформеров. Каждый слой включает два подслоя: механизм внимания (self-attention) и полносвязный нейронный слой с применением функции активации (например, ReLU). Структура трансформера позволяет моделировать контекстуальные зависимости между словами в предложении.
- **Пулинг**: для получения фиксированного размера внутреннего векторного представления всего предложения из выходов трансформеров, BERT использует пулинг. Обычно используется пулинг на основе среднего значения, который вычисляет среднее значение всех выходов токенов.

3. Методы семантического сравнения текстов с помощью BERT

В рамках исследования, основанного на использовании алгоритмов машинного обучения для семантического сравнения текстов, предполагается проведение ряда процедур, основанных на использовании нейросетевых моделей. Главная цель этих действий заключается в реализации алгоритма, способного сравнивать два текстовых фрагмента и оценивать степень их семантической близости.

В работе рассматриваются два текста: исходный текст и текст, с которым проводится сравнение на схожесть. Эти два фрагмента подаются на вход двум одинаковым нейронным сетям, основанным на архитектуре BERT, и обученным с одинаковыми весами. Это позволяет получить два вектора, отражающих семантические характеристики каждого из текстов. Для решения этой задачи в работе предлагаются два подхода, базирующиеся на использовании передовых нейронных сетей, таких как трансформеры и архитектура BERT.

В рамках исследования предлагаются два подхода к дальнейшему использованию полученных векторов. В рамках первого подхода, считается косинусное или евклидово расстояние между двумя векторами, что дает оценку их семантической близости [3]. Затем обучается дополнительная нейронная сеть, которая интерпретирует полученные расстояния и делает более точные выводы о семантической близости текстов.

Второй подход предполагает обучение нейронной сети непосредственно на основе векторов, полученных от нейросетевых моделей BERT [3]. В этом случае нет необходимости в расчете промежуточных расстояний между векторами. Сеть обучается распознавать и сравнивать вектора напрямую, что позволяет ей более точно оценить степень их семантической близости.

Таким образом, в работе рассматриваются два подхода к решению задачи семантического сравнения текстов, каждый из которых имеет свои преимущества и может быть применим в разных условиях и для решения разных задач. Схема общих возможных решений показаны на рисунке 3.1:

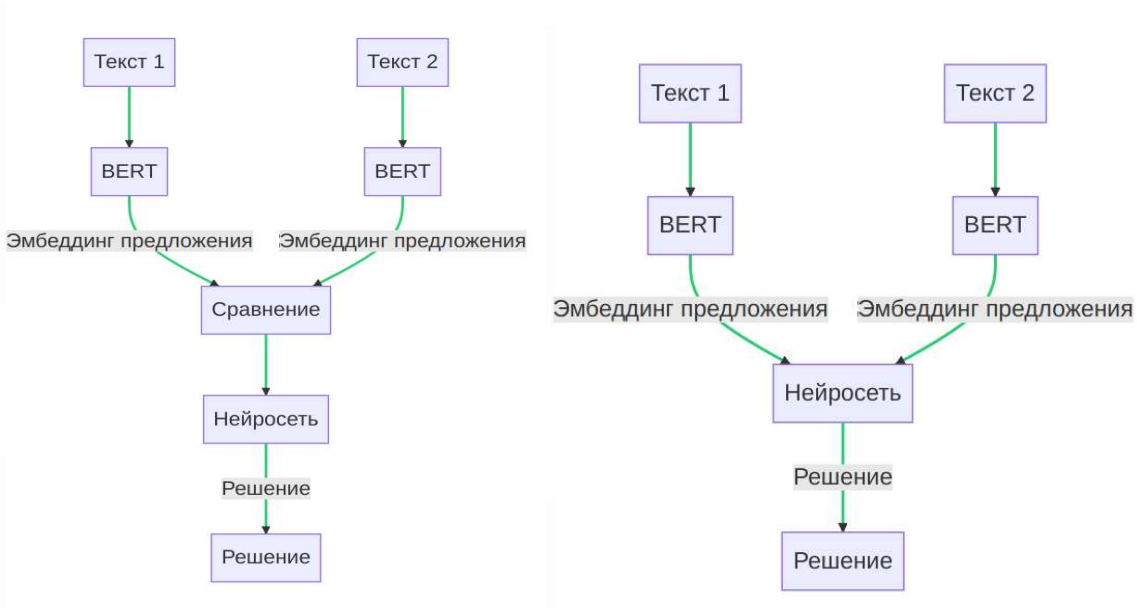

Рисунок 3.1. Схемы возможных решений.

3.1. Сравнение по косинусным расстояниям

Косинусное расстояние является популярной метрикой для сравнения векторов в многомерном пространстве. Оно измеряет косинус угла между двумя векторами, что позволяет оценить их сходство независимо от их длины. В контексте семантического сравнения текстов, косинусное расстояние может быть использовано для сравнения векторных представлений текстов [4].

BERT создает векторные представления текстов, обучаясь на большом корпусе текстов и изучая контекст и семантику слов. Каждое слово в тексте представлено вектором, который отражает его смысл в контексте всего текста. Это делает BERT особенно полезным для семантического сравнения текстов, поскольку его векторные представления учитывают семантические отношения между словами.

Сравнение текстов по косинусным расстояниям с помощью BERT может быть выполнено следующим образом. Сначала каждый текст преобразуется в векторное представление с помощью BERT. Затем вычисляется косинусное расстояние между векторами двух текстов. Меньшее значение косинусного расстояния указывает на большее сходство между текстами.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Этот подход может быть применен в ряде задач, включая определение семантической эквивалентности предложений, классификацию текстов и ранжирование документов. Он также может быть использован для поиска документов, которые семантически близки к заданному запросу, или для группировки документов по темам на основе их семантического сходства.

3.2. Сравнение по евклидовым расстояниям

Евклидово расстояние является одной из наиболее известных метрик для измерения расстояния между двумя точками в многомерном пространстве. В контексте семантического сравнения текстов, евклидово расстояние может быть использовано для сравнения векторных представлений текстов, созданных с помощью модели BERT.

BERT обучается на большом корпусе текстов и создает векторное представление для каждого слова в тексте. Эти векторные представления затем могут быть использованы для семантического сравнения текстов. Каждый текст преобразуется в векторное представление с помощью BERT, а затем вычисляется евклидово расстояние между векторами двух текстов. Меньшее значение евклидового расстояния указывает на большее сходство между текстами.

Этот подход может быть применен в ряде задач, включая определение семантической эквивалентности предложений, классификацию текстов и ранжирование документов. Он также может быть использован для поиска документов, которые семантически близки к заданному запросу, или для группировки документов по темам на основе их семантического сходства.

Однако перед тем, как измерять евклидово расстояние, необходимо нормализовать вектора к единичному вектору, как показано на рисунке 3.2:

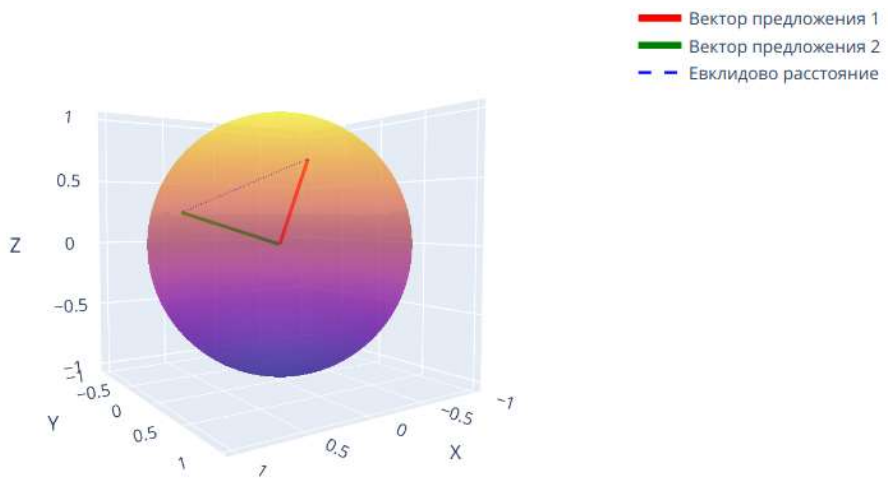

Рисунок 3.2. Визуализация евклидова расстояния для нормализованных векторов.

3.3. Сравнение с помощью обученной нейронной сети по векторам

Сравнение текстов с помощью обученной нейронной сети по векторам представляет собой продвинутый подход, который использует мощь машинного обучения для извлечения более глубоких и точных семантических представлений текста. В этом контексте, архитектура BERT может быть использована для создания векторных представлений текстов, которые затем могут быть использованы как входные данные для нейронной сети.

BERT обучается на большом корпусе текстов и создает векторное представление для каждого слова в тексте. Эти векторные представления затем могут быть использованы для семантического сравнения текстов. Однако, вместо прямого сравнения векторов с помощью метрик, таких как косинусное расстояние, мы можем использовать эти векторы как входные данные для обученной нейронной сети.

Нейронная сеть может быть обучена на задаче сравнения текстов, такой как определение семантической эквивалентности предложений или классификация текстов. В этом случае, нейронная сеть будет использовать векторные представления BERT для изучения более сложных семантических отношений между текстами и для создания более точных предсказаний.

Применение этого подхода может быть очень разнообразным. Он может быть использован для создания системы рекомендаций, которая предлагает пользователям контент, семантически близкий к тому, что они уже просмотрели, или для создания системы поиска, которая возвращает документы, семантически близкие к заданному запросу. Он также может быть использован для автоматической классификации документов по темам или для определения степени сходства между двумя текстами.

4. Реализация алгоритма машинного обучения для семантического сравнения текстов

4.1. Выбор датасета

4.1.1. О датасете

STS Benchmark (Semantic Textual Similarity Benchmark) [5] — это широко используемый набор данных и метрик для оценки и сравнения алгоритмов семантического сравнения текстов. Он предназначен для измерения семантической близости или сходства между парами предложений.

STS Benchmark состоит из большого набора пар предложений, для каждой из которых имеется асессорская оценка семантического сходства в виде числового значения. Оценки задаются на шкале от 0 до 5, где 0 означает полное отсутствие семантической близости, а 5 означает полное семантическое сходство.

Набор данных STS Benchmark включает разнообразные типы пар предложений, включая пары синонимов, пары сходных предложений с небольшими вариациями, а также пары предложений с разными стилями и контекстами. Это позволяет оценить производительность алгоритмов семантического сравнения текстов в различных сценариях.

Для оценки алгоритмов семантического сравнения текстов на наборе данных STS Benchmark используются различные метрики, такие как корреляция Пирсона и коэффициент Спирмена. Эти метрики измеряют соответствие оценок, полученных алгоритмом, с оценками асессоров.

STS Benchmark является важным инструментом для сравнения и оценки алгоритмов семантического сравнения текстов. Он позволяет исследователям и разработчикам оценить производительность своих моделей на широком наборе данных и сравнить их с результатами других подходов. Это способствует развитию и улучшению методов семантического сравнения текстов и продвижению в области обработки естественного языка.

4.1.2. Существующие альтернативы

В дополнение к STS Benchmark, существуют другие бенчмарки и метрики для оценки и сравнения алгоритмов семантического сравнения текстов. Ниже перечислены некоторые из них:

Paraphrase Database (PPDB): PPDB [6] — это база данных семантически эквивалентных и парафразных предложений. Она содержит большой набор пар предложений с различными уровнями парафраза. PPDB может использоваться для оценки алгоритмов семантического сравнения текстов, особенно в контексте поиска и аугментации парафраза.

SICK (Sentences Involving Compositional Knowledge) Dataset [7]: SICK Dataset является еще одним набором данных для семантического сравнения текстов. Он включает пары предложений с асессорскими оценками и различными семантическими отношениями, такими как парафраз, противоположность и несоответствие.

MSR (Microsoft Research) Paraphrase Corpus [8]: это набор данных, разработанный Microsoft Research, состоящий из пар предложений, которые либо являются парафразами, либо не являются ими. Он предоставляет базовый набор для оценки алгоритмов семантического сравнения текстов и задачи определения парафраза.

MAP метрика (Mean Average Precision) [9]: MAP является распространенной метрикой для оценки точности ранжирования в задаче семантического сравнения текстов. Она учитывает не только правильные ранжирования пар предложений, но и положение правильного ранжирования в списке рекомендаций.

4.1.3. Структура набора данных

Данный набор данных содержит два предложения и степень их сходства в пределах от нуля до 5. Примеры из набора данных показаны на рисунке 4.1.

Первое предложение	Второе предложение	Степень сходства
A man is cutting an onion.	A man cuts an onion.	5
A man is cycling.	A man is talking.	0.6
A man is singing while playing the guitar.	A man is playing a guitar.	3.6

Рисунок 4.1. Структура набор данных.

4.2. Описание процесса подготовки данных

Данные для обучения необходимо нормализовать. Для нормализации значений в пределах от нуля до единицы для обучения нейронных сетей необходимо определить минимальное и максимальное значение каждого признака в наборе данных. Затем применить формулу: normalized_x = (x - min) / (max - min) для каждого значения признака x. Это помогает упростить обучение, предотвратить доминирование признаков и улучшить обобщающую способность модели.

Также для первой версии были переведены предложения с английского на русский язык. Но в дальнейшем пришлось отказаться от этого решения, потому что результаты оказались хуже, чем без использования перевода. На рисунке 4.2 показано несколько примеров из набора данных, переведенных на русский язык.

Первое предложение	Второе предложение	Степень сходства
Мужчина режет лук.	Человек разрезает лук.	1.0
Мужчина едет на велосипеде.	Человек говорит.	0.12
Человек поет во время игры на гитаре.	Мужчина играет на гитаре.	0.72

Рисунок 4.2. Набор данных после перевода и нормализации меток.

4.3. Описание выбранного подхода и его обоснование

В ходе работы будет реализовано несколько вариантов. Самые первые эксперименты проводились с использованием сравнения косинусных расстояний и предобученной модели RuBERT [10] для извлечения внутренних векторных представлений.

RuBERT (Russian BERT) — это модель глубокого обучения, основанная на архитектуре BERT (Bidirectional Encoder Representations from Transformers)

[10], специально разработанная для обработки русскоязычного текста. BERT является предварительно обученной моделью, которая способна эффективно понимать семантику и контекст в текстовых данных.

Затем были попытки улучшить алгоритм путем изменения базовой модели BERT. Был проведен сравнительный анализ моделей. И выбрана наиболее оптимальная из них по критерию скорости работы и качеству.

5. Эксперименты и результаты

5.1. Первые попытки построить алгоритм

Для первых экспериментов был использован набор данных с переведенными на русский язык предложениями и моделью RuBERT для получения внутреннего векторного представления.

5.1.1. Получение внутреннего векторного представления

Внутренние векторные представления брались с самого последнего слоя нейронной сети BERT, а затем сравнивались по косинусным расстояниям.

Для общей оценки корректности работы модели с данными было принято решение изучить диаграмму распределения целевой метки от косинусного расстояния, а также рассчитать корреляцию, полученную данным путем. На рисунке 5.1 представлена диаграмма рассеяния метки от косинусного расстояния.

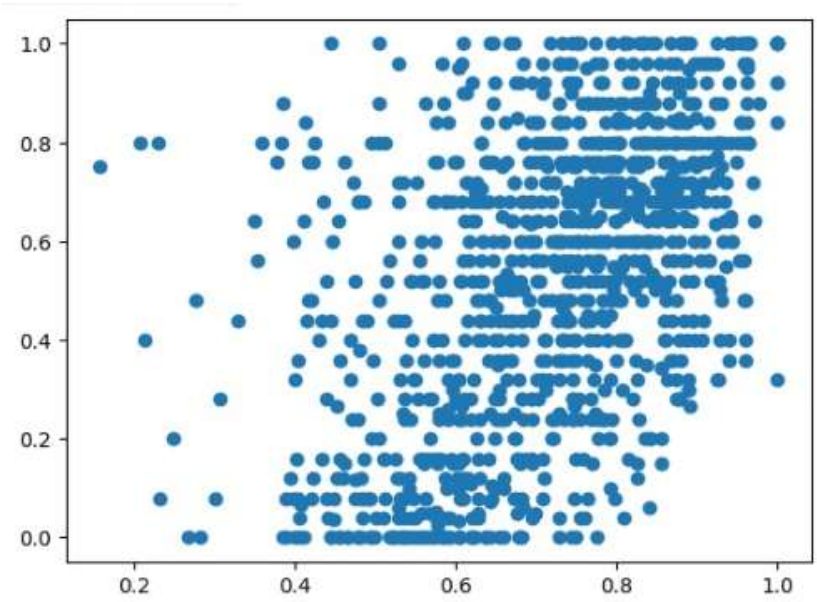

Рисунок 5.1. Диаграмма рассеяния метки от косинусного расстояния.

Значение корреляции между целевой меткой и косинусным расстоянием составило 0.484, что не выглядит хорошим показателем, поскольку современные алгоритмы имеют этот показатель в пределах от 0.799 до 0.927. Тем не менее, было принято решение провести обучение нейронной сети и оценить полученные результаты.

5.1.2. Построение нейронной сети для получения результата

Было принято решение разработать нейронную сеть, которая бы определяла семантическую близость на основе евклидовых расстояний. Такой подход позволяет измерять расстояние между векторами представлений и использовать его в качестве метрики для определения степени семантической близости текстов. График потерь можно рассмотреть на рисунке 5.2.

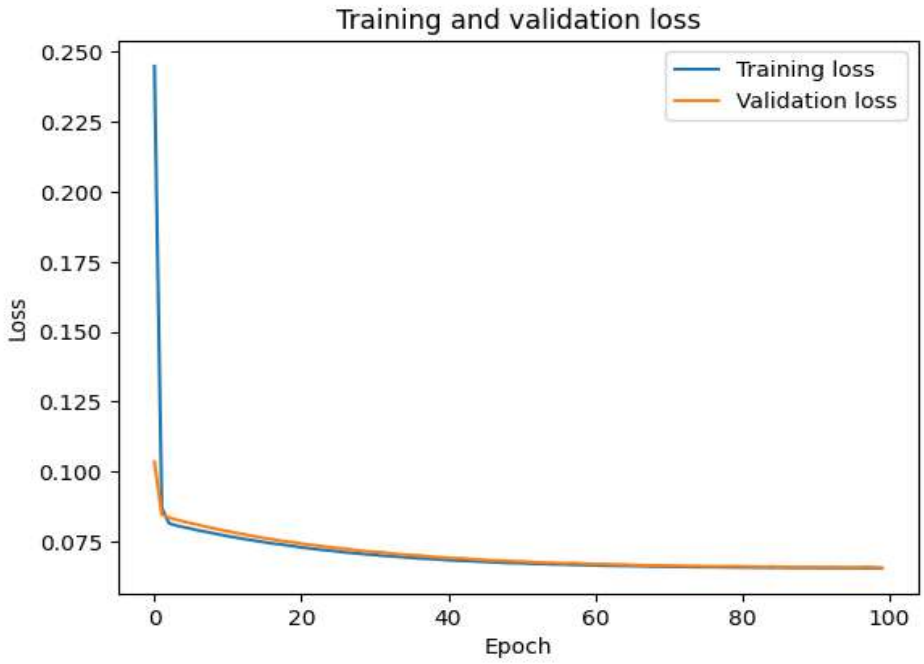

Рисунок 5.2. График потерь на тестовой и обучающей выборках при обучении НС на косинусных расстояниях

Как видно на графике, нейронная сеть обучается и делает это без переобучения, поскольку среднеквадратичная ошибка уменьшается на обоих выборках: на тестовой и на тренировочной. Качество полученного алгоритма можно измерить по полученной итоговой среднеквадратичной ошибке, которая составила 0,0656. Данная нейронная сеть позволяет лучше интерпретировать полученные расстояния в качестве искомых меток.

5.2. Попытки улучшить модель

Далее было найдено несколько способов улучшить модель:

- протестировать альтернативные модели BERT;
- изменить язык на оригинальный (английский);
- конвертировать косинусные расстояния в евклидовы;

Далее все эксперименты проводились с набором данных с оригинальным языком.

5.2.1. Сравнительный анализ предобученных моделей BERT

Для улучшения модели можно начать с замены модели, используемой для получения внутренних векторных представлений текстов. В таблице 5.1 для сравнения представлены альтернативные модели.

Таблица 5.1. Альтернативные модели для сравнения

Обозначение модели	Название модели
m1	distilbert-base-multilingual-cased [11]
m2	sentence-transformers/distiluse-base-multilingual-cased-v1 [12]
m3	inkoziev/sbert_synonymy [13]
m4	sentence-transformers/stsb-xlm-r-multilingual [14]
m5	paraphrase-multilingual-MiniLM-L12-v2 [15]

После получения внутренних представлений, можно приступить к оценке модели на основе нескольких критериев.

Первым критерием может быть корреляция между полученными векторами предложений и метками набора данных. Это означает, что модель должна создавать внутренние векторные представления, которые имеют высокую корреляцию с истинными значениями или оценками в наборе данных. Чем выше корреляция, тем более точно модель учитывает семантическую близость предложений.

Вторым критерием может быть величина полученных внутренних векторных представлений текстов. Хорошая модель должна производить

внутренние векторные представления, которые сохраняют важные семантические свойства текстов, при этом имея умеренный размер. Слишком большие вектора могут привести к увеличению вычислительной сложности и потреблению памяти.

Оценивая модели по этим критериям, можно выбрать лучшую модель для задачи семантического сравнения текстов. Это позволит улучшить качество представлений текстов и повысить точность в определении их семантической близости. В таблице 5.2 приведены результаты сравнения моделей.

Таблица 5.2. Результаты сравнения моделей

	m1	m2	m3	m4	m5
Значение корреляции с меткой	-0.570	-0.822	-0.635	-0.866	-0.86
Время вычисления 2000 эмбеддингов (с.)	19.81	21.26	9.42	38.44	34.41

При проведении сравнительного анализа моделей для семантического сравнения текстов было выяснено, что две модели показали наилучшие результаты. Одной из них является **paraphrase-multilingual-MiniLM-L12-v2**, а другой - **sentence-transformers/stsb-xlm-r-multilingual**.

Модель m5 - paraphrase-multilingual-MiniLM-L12-v2 была выбрана в качестве оптимального варианта несмотря на то, что ее корреляция была немного ниже по сравнению с sentence-transformers/stsb-xlm-r-multilingual, имеет более низкую размерность внутреннего векторного представления - 384. Это позволяет повысить скорость вычислений, что является значимым преимуществом в сфере обработки больших объемов текстовых данных. На рисунке 5.3 представлена диаграмма рассеяния косинусных расстояний от целевой метки.

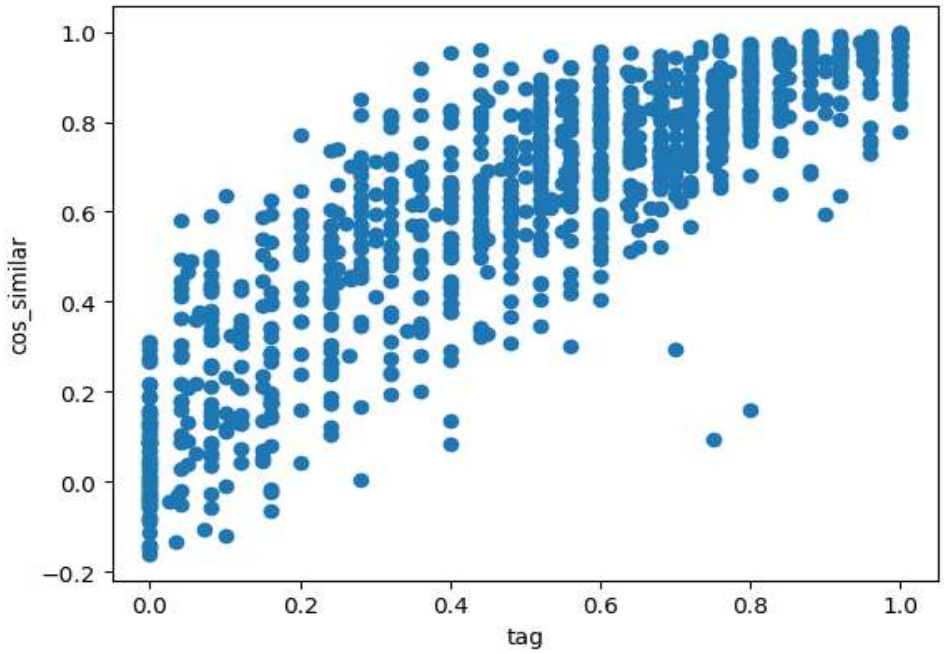

Рисунок 5.3. Диаграмма рассеяния косинусных расстояний от целевой метки

В результате экспериментов было установлено, что корреляция между косинусным расстоянием, основанным на внутренних векторных представлениях модели **paraphrase-multilingual-MiniLM-L12-v2**, и целевой переменной составляет 0,86. Это указывает на сильную связь между семантической близостью текстов, определенной моделью, и истинными значениями в целевой переменной. Такая высокая корреляция подтверждает эффективность модели в задаче семантического сравнения текстов.

Таким образом, модель **paraphrase-multilingual-MiniLM-L12-v2** представляет собой перспективный вариант для решения задачи семантического сравнения текстов, благодаря своей относительно низкой размерности векторов внутреннего представления текстов и высокой корреляции с целевой переменной.

5.2.2. Преобразование косинусных расстояний в евклидовы

При проведении эксперимента с использованием косинусного расстояния для измерения семантической близости текстов было замечено, что полученный график зависимости косинусного расстояния от целевой переменной может быть неудовлетворительным.

Для улучшения результатов, возникла идея нормализовать векторы и использовать евклидову метрику вместо косинусного расстояния. Нормализация векторов проводилась путем приведения их к единичной длине. На рисунке 5.4 представлена диаграмма рассеяния нормализованных евклидовых расстояний к единичной длине от целевой метки

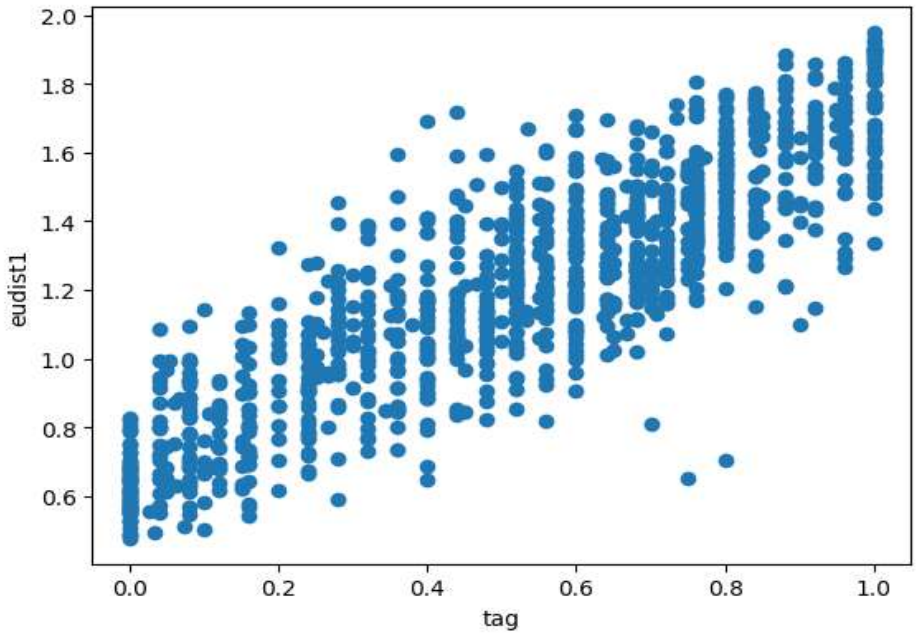

Рисунок 5.4. Диаграмма рассеяния нормализованных к единичной длине евклидовых расстояний от целевой метки

После применения нормализации векторов и перехода к евклидовой метрике, было обнаружено, что значение корреляции улучшилось по сравнению с предыдущим подходом: значение составило 0.869 вместо 0.86 при косинусных расстояниях. Кроме того, график зависимости евклидовой метрики от целевой переменной оказался более удовлетворительным. На рисунке 10 приведена матрица корреляций.

	tag	cos_dist	cos_similar	eudist1
tag	1.000000	-0.860156	0.860156	-0.868712
cos_dist	-0.860156	1.000000	-1.000000	0.974745
cos_similar	0.860156	-1.000000	1.000000	-0.974745
eudist1	-0.868712	0.974745	-0.974745	1.000000

Рисунок 10. Матрица корреляций

Таким образом, использование евклидовой метрики вместо косинусного расстояния позволило достичь более высокого значения корреляции и получить более убедительный и наглядный график зависимости между метрикой и целевой переменной.

Это подтверждает эффективность использования евклидовой метрики для задачи семантического сравнения текстов и может являться предпочтительным вариантом.

5.2.3. Построение нейронной сети на основе внутренних векторных представлений

При решении задачи семантического сравнения текстов возникла необходимость выбрать один из двух подходов. Первый путь заключается в обучении нейронной сети на сокращенных векторах, полученных с использованием методов уменьшения размерности. Второй путь предполагает обучение нейронной сети на расстояниях между векторами предложений, чтобы определить степень их схожести.

Рассмотрим обучение нейронной сети по полученным векторам. Для снижения размерности векторов был применен метод главных компонент (PCA). С помощью этого метода размерность векторов была уменьшена до 30.

Нейронная сеть была обучена на полученных векторах. При этом использовались различные топологии, гиперпараметры, функции активации и методы регуляризации. Топологии нейронных сетей варьировались от простых однослойных моделей до более сложных глубоких архитектур. ReLU была выбрана как функция активации. Регуляризация была выбрана комбинированная в виде L1 или L2 регуляризации, которая использовалась для уменьшения переобучения.

По результатам обучения нейронной сети по векторам получен относительно высокий показатель среднеквадратичной ошибки, который составляет 0.05. Это свидетельствует не об очень хорошей способности модели обобщать и предсказывать семантическую близость текстов.

5.2.4. Построение нейронной сети по евклидовым расстояния

Альтернативным подходом было обучение нейронной сети по евклидовым расстояниям между парами векторов предложений.

Аналогично предыдущему случаю, были исследованы различные гиперпараметры, включая топологию нейронной сети, функции активации и регуляризацию.

В результаты была выбрана простая нейронная сеть с двумя скрытыми слоями, которые содержат 5 и 4 нейрона в слоях соответственно. Функция активации ReLU показала себя лучше всего. Методы регуляризации не потребовались. В качестве оптимизатора были выбраны моменты Нестерова.

По результатам обучения нейронной сети в течение 1000 эпох по евклидовым расстояниям получено значение среднеквадратичной ошибки на тестовой выборке, который составляет примерно 0.019. Это значительно меньшее значение указывает на то, что модель успешно обучается предсказывать схожесть предложений на основе их евклидовых расстояний.

Несмотря на то, что оба значения (0.05 и 0.019) кажутся довольно малыми, разница между ними имеет большое значение при оценке производительности моделей. Метрика среднеквадратичной ошибки (MSE) усиливает влияние больших ошибок за счет квадратичной зависимости, что делает ее чувствительной даже к небольшим изменениям. Именно поэтому даже небольшое различие в MSE может иметь значительное влияние на качество прогнозов модели.

Таким образом, модель, обученная на основе евклидовых расстояний, предсказывает семантическую близость текстов значительно точнее, чем модель, обученная по векторам и тем более, чем первоначальная модель, которая была основана на модели RuBERT с сравнением косинусных расстояний с среднеквадратичной ошибкой равной 0.0656.

5.3. Примеры работы нейронной сети

Далее в таблице 5.3 приведены примеры работы нейронной сети на разных текстах из набора данных:

Таблица 5.3. Примеры работы нейронной сети на исходном наборе данных.

Текст 1	Текст 2	Оценка нейронной сети	Верный ответ
Мужчина режет лук	Человек разрезает лук	0.797	1
Мужчина едет на велосипеде	Человек говорит	0.089	0.12
Мужчина играет на гитаре	Человек поёт во время игры на гитаре	0.722	0.72

Нейронная сеть, обученная на евклидовых расстояниях, эффективно определяет семантическую близость текстов. Это подтверждается высокой степенью совпадения между оценками, выданными нейронной сетью, и реальными значениями.

В примерах, приведенных в таблице 3, оценки сети очень близки к истинным значениям. В особенности, модель удачно оценивает как близкие по смыслу предложения ("Мужчина режет лук" и "Человек разрезает лук"), так и далекие ("Мужчина едет на велосипеде" и "Человек говорит").

Отметим, что примеры из таблицы 3 представляют собой короткие тексты, которые часто представлены в типичных наборах данных для семантического сравнения текстов. Было бы полезно проверить работу алгоритма на больших текстах или на данных, имеющих большее семантическое разнообразие, чтобы более полно оценить его эффективность.

Также необходимо отметить, что среднеквадратичная ошибка между оценками модели и истинными значениями оказалась значительно ниже, чем при использовании других подходов, что говорит о высокой эффективности применяемого подхода.

Таким образом, анализ данных из таблиц указывает на успешность применения обученной нейронной сети для решения задачи семантического сравнения текстов. Однако для более уверенного вывода следует провести дополнительные тесты, включая применение на больших текстах и на различных типах данных.

В таблице 5.4 приведены примеры работы нейронной сети на разных текстах большей длины:

Таблица 5.4. Примеры работы нейронной сети на больших текстах.

Текст 1	Текст 2	Оценка нейронной сети	Примечание
Рассвет пролил свет на горный пейзаж, открывая величественные снежные вершины и мирно стоящие леса. Над горами парил орел, знак могущества и свободы, его глаза следили за каждым движением внизу. Земля была еще тепла от вчерашнего дня, и эти последние остатки тепла смешались с прохладным утренним воздухом, создавая ощущение комфорта и умиротворения.	Утренний свет раскрыл горизонты горной панорамы, выявляя внушительные заснеженные пики и неподвижные лесные массивы. Над высотами парил коршун, символ величия и независимости, его зоркие глаза наблюдали за всем происходящим внизу. Почва еще сохраняла тепло предыдущего дня, и эти последние крупицы тепла сочетались с прохладной утренней атмосферой, пробуждая чувство уюта и спокойствия.	0.873	Тексты несут одинаковый смысл, но совершенно разными речевыми конструкциями.
Космическая техника и исследования совершают революцию в нашем понимании Вселенной. От микроскопических частиц до галактик, наука продолжает расширять наши знания о том, как работает Вселенная. С помощью телескопов, спутников и межпланетных зондов мы можем изучать далекие звезды и планеты, а также пытаться ответить на вопросы о происхождении и будущем космоса.	Балет — это восхитительная форма искусства, сочетающая грацию, движение и музыку. Он требует от исполнителей не только физической силы и гибкости, но и эмоционального выражения. Балетные спектакли могут перенести зрителей в разные эпохи и культуры, предлагая новые перспективы и понимание мира. С самых первых шагов на пуантах до последнего поклона, балет остается удивительным зрелищем для любого возраста.	0.1711	Тексты несхожие.
Автомобильная индустрия претерпевает революцию благодаря электрическим транспортным средствам. Благодаря их низкому уровню выбросов и возможности	Автономные автомобили становятся все более распространенными и переопределяют подходы к транспортным системам. Эти транспортные средства оснащены современными	0.5121	Пример работы алгоритма на текстах на похожие темы, но о разных вещах.

использования возобновляемых источников энергии, электромобили представляют собой привлекательный вариант для устойчивого будущего. Многие автомобильные компании уже инвестируют миллиарды в разработку и производство электромобилей, обеспечивая быстрый рост этого рынка. Тем не менее, остаются вызовы, такие как создание эффективной инфраструктуры зарядки и улучшение технологии аккумуляторов.	технологиями, позволяющими им самостоятельно воспринимать окружающую среду и принимать решения о передвижении без непосредственного управления водителем. Использование автономных автомобилей может улучшить безопасность дорожного движения, уменьшить пробки и облегчить перемещение для людей, которые не могут или не хотят водить автомобиль. Однако развитие этой технологии также вызывает вопросы в области регулирования, этики и безопасности.		

Алгоритм успешно определяет семантическую близость текстов, которые содержат одинаковый смысл, но выражены разными речевыми конструкциями. Это подтверждается первым примером, где оценка нейронной сети составляет 0.873 для двух очень схожих текстов, но с различной формулировкой.

Алгоритм также эффективен в различении текстов, которые не имеют семантической связи. Во втором примере, где тексты описывают две разные темы (космос и балет), нейронная сеть правильно приписала низкую оценку семантической близости (0.1711), что соответствует истинному семантическому расстоянию между ними.

В третьем примере, где оба текста связаны с автомобильной индустрией, но рассматривают разные аспекты (электромобили и автономные автомобили), алгоритм дал среднюю оценку семантической близости (0.5121). Это указывает на то, что алгоритм может определять, что тематика текстов схожая, но суть у них разная.

В целом, алгоритм показывает хорошую производительность на представленных примерах.

5.4. Вывод по результатам разработки алгоритма

В рамках этого исследования проводился сравнительный анализ разнообразных моделей и подходов к семантическому сравнению текстов, включая использование внутренних векторных представлений текста и применение нейронных сетей для предсказания степени их семантической близости.

В начале использовались несколько предобученных моделей, включая paraphrase-multilingual-MiniLM-L12-v2 и sentence-transformers/stsb-xlm-r-multilingual, для получения векторных представлений текстов. Из этого списка paraphrase-multilingual-MiniLM-L12-v2 была выбрана в силу относительно низкой размерности внутренних векторных представлений и высокой корреляции с целевой переменной.

Затем было применено два подхода к обучению нейронных сетей для решения задачи: обучение на уменьшенных векторах с использованием метода главных компонент и обучение на евклидовых расстояниях между парами векторов предложений. Первый подход показал результат среднеквадратичной ошибки 0.05, в то время как второй подход, основанный на евклидовых расстояниях, показал существенно более точные прогнозы с среднеквадратичной ошибкой 0.019.

Таким образом, результаты этого исследования подтверждают эффективность использования евклидовой метрики для семантического сравнения текстов и применения нейронных сетей для обучения на основе этих расстояний. Эти выводы открывают новые перспективы для улучшения качества моделей обработки естественного языка и расширения их практического применения.

6. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

6.1.Предпроектный анализ

6.1.1 Потенциальные потребители результатов исследования

Целью проекта является разработка нейросетевого алгоритма для семантического сравнения текстов. Данная работа может проводить сравнение двух текстовых блоков по их смыслу. Такой работой могут заинтересоваться, к примеру, системы для информационного поиска, рекомендательные системы, системы категоризации текстов.

Сегментировать рынок можно по степени потребности использования возможностей системы. В таблице 6.1 показана сегментация рынка.

Таблица 6.1 – Карта сегментирования рынка

		Вид	
		Улучшение систем релевантного поиска	Анализ текстовых данных
Размер	Крупные		
	Средние		
	Малые		

	Поисковые системы		Исследователи и аналитики		Рекомендательные системы

6.1.2 Анализ конкурентных технических решений с позиции ресурсоэффективности и ресурсосбережения

Задача имеет большое количество решений, большинство из которых предоставляются платно, как сервис. Выделим 2 популярных конкурентных решения.

1. **Copyleaks** – это приложение для обнаружения плагиата на базе искусственного интеллекта, которое выходит за рамки дословного сопоставления и идентифицирует перефразированный и похожий по смыслу текст.

2. **P2pi.ru** – онлайн инструмент для сравнения двух текстов на процентное совпадение содержания. Сервис создан для копирайтинга и рерайта с целью объективно сравнить содержание первоисточника и обработанного текста.

Таблица 6.2 – Оценочная карта для сравнения конкурентных решений

Критерии оценки	Вес критерия	Баллы			Конкурентоспособность		
		$Б_ф$	$Б_{к1}$	$Б_{к2}$	$К_ф$	$К_{к1}$	$К_{к2}$
1	2	3	4	5	7	8	9
Технические критерии оценки ресурсоэффективности							
1. Повышение производительности труда	0,078	5	5	2	0,39	0,39	0,156
2. Простота эксплуатации	0,055	4	4	3	0,22	0,22	0,165
3. Интерпретируемость	0,070	3	5	5	0,21	0,35	0,35
4. Точность	0,078	3	5	1	0,23	0,39	0,078
5. Удобство эксплуатации	0,055	4	4	3	0,22	0,22	0,165
6. Степень автоматизации	0,055	5	5	5	0,28	0,28	0,275
7. Энергоэкономичность	0,016	2	4	4	0,03	0,06	0,064
8. Требования к оборудованию	0,039	2	3	5	0,08	0,12	0,195
9. Работа по сети Интернет	0,008	5	5	5	0,04	0,04	0,04
10. Дополнительные функциональные возможности	0,055	1	4	1	0,06	0,22	0,055
Экономические показатели эффективности							
1. Конкурентоспособность продукта	0,078	3	4	1	0,23	0,31	0,078
2. Цена	0,070	2	4	2	0,14	0,28	0,14
3. Срок выхода на рынок	0,063	2	4	5	0,13	0,25	0,315
4. Скорость обучения использованию	0,063	5	5	5	0,32	0,32	0,315
5. Стоимость масштабирования	0,063	3	3	3	0,19	0,19	0,189
6. Стоимость модификации	0,063	3	3	3	0,19	0,19	0,189
7. Стоимость внедрения	0,078	5	3	2	0,39	0,23	0,156
8. Уровень проникновения на рынок	0,016	1	4	2	0,02	0,06	0,032
Итого	1	58	74	57	3,35	4,12	2,957

6.1.2 SWOT-анализ

В ходе SWOT-анализа были определены приоритетные направления развития разработки, а также способы для нивелирования слабых сторон разработки и противодействия возможным угрозам. Это позволяет перейти ко второму этапу анализа.

Для второго этапа анализа необходима интерактивная матрица проекта, которая представлена в таблице 6.3.

Таблица 6.3 – SWOT-анализ

	Сильные стороны С1. Использование передовых методов машинного обучения, таких как трансформеры и BERT, для семантического сравнения текстов. С2. Возможность анализа больших объемов текстовых данных. С3. Повышение производительности и эффективности сравнения текстов. С4. Потенциал для улучшения систем поиска и рекомендаций на интернет-платформах.	**Слабые стороны** Сл1. Потребность в больших объемах данных для обучения модели. Сл2. Возможные сложности в интерпретации результатов семантического сравнения. Сл3. Потребность в высокой вычислительной мощности для работы алгоритма.
Возможности В1. Растущий спрос на интеллектуальный анализ текстовых данных в различных отраслях. В2. Возможность применения в областях, требующих семантического анализа текста, таких как маркетинг, социальные науки, искусственный интеллект и другие. В3. Потенциал для дальнейшего развития и улучшения алгоритмов семантического сравнения текстов.	**В1 + С1:** Растущий спрос на анализ текстовых данных увеличивает значение использования передовых методов машинного обучения для семантического сравнения текстов. **В2 + С2:** Применение методов в областях, требующих семантического анализа текста, подразумевает обработку больших объемов текстовых данных. **В3 + С4:** Потенциал для дальнейшего развития и улучшения алгоритмов обещает улучшение систем поиска и рекомендаций на интернет-платформах.	**В1 + Сл1:** Необходимость больших объемов данных для обучения модели может представлять сложности в условиях растущего спроса на интеллектуальный анализ текстовых данных. **В2 + Сл2:** Возможные сложности в интерпретации результатов семантического сравнения могут оказывать влияние на применение методов в различных областях. **В3 + Сл3:** Потенциал для дальнейшего развития и улучшения алгоритмов требует высокой вычислительной мощности для работы алгоритма.
Угрозы У1. Конкуренция со стороны других передовых методов	**У1 + С1:** Передовые методы машинного обучения, такие как трансформеры и BERT, могут	**У1 + Сл1:** Конкуренция может усилиться в условиях необходимости больших

машинного обучения и нейронных сетей. У2. Возможные проблемы с защитой данных и конфиденциальностью при работе с текстовыми данными. У3. Технические сложности и проблемы с производительностью, связанные с обработкой больших объемов текстовых данных.	привлечь конкурентов, что увеличивает значение использования этих методов для поддержания конкурентоспособности. **У2 + С2:** Обработка больших объемов текстовых данных может подвергать модель риску нарушения защиты данных и конфиденциальности. **У3 + С4:** Потенциал для улучшения систем поиска и рекомендаций на интернет-платформах подвержен техническим сложностям и проблемам с производительностью.	объемов данных для обучения модели. **У2 + Сл2:** Возможные сложности в интерпретации результатов семантического сравнения могут увеличивать риски, связанные с защитой данных и конфиденциальностью. **У3 + Сл3:** Требование высокой вычислительной мощности для работы алгоритма подвержено техническим сложностям и проблемам с производительностью.

Таким образом, проведенный SWOT-анализ наглядно демонстрирует преимущество сильных сторон данного проекта. Анализ отражает возможности, которые возникают из-за наличия сильных сторон и угрозы, которые могут повлиять на эти сильные стороны. Самой сильной стороной проекта является тот факт, что проект использует передовые методы машинного обучения - трансформеры и BERT, что позволяет обрабатывать большие объемы текстовых данных и улучшает качество семантического сравнения текстов. Это делает ваше решение особенно ценным для улучшения систем поиска и рекомендаций на интернет-платформах.

Слабыми сторонами проекта являются несколько факторов: результаты семантического сравнения текстов могут быть сложными для интерпретации, также алгоритм требует высокой вычислительной мощности, что может представлять технические сложности и проблемы с производительностью.

6.1.3 Оценка готовности проекта к коммерциализации

Опишем степень готовности проекта к коммерциализации в таблице 6.4. В этой таблице предоставлен перечень вопросов, позволяющих выяснить проработанность проекта с точки зрения коммерциализации и компетенции разработка.

Таблица 6.4 – Оценка степени готовности проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	4	4
2.	Определены перспективные направления коммерциализации научно-технического задела	3	3
3.	Определены отрасли и технологии (товары, услуги) для предложения на рынке	4	3
4.	Определена товарная форма научно-технического задела для представления на рынок	2	2
5.	Определены авторы и осуществлена охрана их прав	2	2
6.	Проведена оценка стоимости интеллектуальной собственности	2	2
7.	Проведены маркетинговые исследования рынков сбыта	3	2
8.	Разработан бизнес-план коммерциализации научной разработки	2	2
9.	Определены пути продвижения научной разработки на рынок	3	3
10.	Разработана стратегия (форма) реализации научной разработки	3	2
11.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	2	2
12.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	4	3
13.	Проработаны вопросы финансирования коммерциализации научной разработки	3	3
14.	Имеется команда для коммерциализации научной разработки	2	2
15.	Проработан механизм реализации научного проекта	3	3
	ИТОГО БАЛЛОВ	42	38

Согласно таблице 6.4 готовность проекта к коммерциализации находится на среднем уровне. Это уровень можно повысить, более детально проработав бизнес-план и вопросы международного сотрудничества, маркетинговые исследования, а также за счет повышения квалификации разработчика в области интеллектуального права, продвижения продукта и других вопросов.

6.1.4 Методы коммерциализации результатов научно-технического исследования.

Для проекта в качестве наиболее подходящих методов для коммерциализации были выбраны следующие методы:

1. **Передача ноу-хау:** Разработка представляет собой сложную технологию, требующую знаний и опыта для эффективного применения и доработки. В этом контексте передача ноу-хау, то есть предоставление знаний и опыта по использованию технологии другим лицам, может стать эффективным инструментом коммерциализации. Это позволит привлекать клиентов, которые ценят глубину понимания технологии и ее возможностей, а также обеспечит добавочную стоимость к самой технологии.

2. **Организация собственного предприятия:** Исходя из специфики разработки и ее потенциала, организация собственного предприятия для продажи услуг или продуктов на основе данной технологии может быть оправдана. Собственное предприятие обеспечит полный контроль над технологией, ее развитием и коммерческим использованием.

3. **Инжиниринг:** Этот метод коммерциализации включает в себя предоставление комплекса инженерно-технических услуг, связанных с проектированием, внедрением и усовершенствованием технологических процессов на предприятии заказчика. В случае с данной разработкой, можно предложить услуги по внедрению и доработке алгоритмов семантического сравнения текстов для улучшения процессов обработки и анализа больших объемов текстовых данных у клиента. Данный метод особенно эффективен, если клиенты нуждаются в индивидуальных решениях, нацеленных на конкретные задачи или бизнес-процессы.

6.2. Инициация проекта

Устав научного проекта магистерской работы:

1. Цели и результат проекта

Приведем информацию о заинтересованных сторонах в таблице 6.5.

Таблица 6.5 – Заинтересованные стороны проекта

Заинтересованные стороны
1. Разработчик: они будут прямо задействованы в разработке, тестировании и улучшении алгоритма.
2. Пользователи интернет-платформ: они могут получить прямую выгоду от улучшенных систем поиска и рекомендаций, основанных на алгоритме семантического сравнения текстов.
3. Компании, занимающиеся анализом текстовых данных: они могут использовать полученные результаты для улучшения своих продуктов и услуг.

Цели и результат проекта представлены в таблице 6.6.

Таблица 6.6 – Цели и результат проекта

Цели проекта	– Изучить предметную область – Разработка алгоритма машинного обучения для сравнения двух текстов по их смыслу, используя трансформеры и нейронную архитектуру BERT – Улучшение текущих способов сравнения текстов
Ожидаемые результаты	– Разработанный алгоритм, способный сравнивать смысл двух текстов и определять схожесть – Совершенствование навыков в области машинного обучения, искусственного интеллекта и обработки текстовых данных.

2. Организационная структура проекта. Данная работа была инициализирована научным руководителем и магистрантом. В таблице 6.7 Приведена организационная структура проекта.

Таблица 6.7 – Организационная структура проекта

№	ФИО, Основное место работы, Должность	Роль в проекте	Функции	Трудозатраты, дни.
1	Цыденов С.Б. студент	Исполнитель проекта	Исследование предметной области Проектирование, реализация и тестирование алгоритма Написание отчета	121
2	Доцент ОИТ ИШИТР, Спицын В.Г., д.т.н	Руководитель проекта	Консультирование по предметной области	5
Итого				126

6.3. Планирование управления научно-техническим проектом

6.3.1 План проекта

Для организации процесса разработки был определен ряд задач для каждого этапа работы. Для реализации проекта необходимо 2 исполнителя: Научный руководитель (НР), инженер(И).

Таблица 6.8 – Календарный план-график проведения НИОКР по теме

№	Вид	Исполнитель	Т рабочих	Продолжительность														
				Янв.			Февр.			Март			Апр.			Май		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	Выбор научного руководителя работы	С	8															
2	Составление и утверждение темы работы	С	4															
		Р	2															
3	Составление календарного плана-графика выполнения работы	С	5															
		Р	1															
4	Подбор и изучение литературы по	С	18															

	теме работы																	
5	Анализ предметной области	С	14															
6	Проверка гипотезы прогнозирования по фундаментальным показателям	С	8															
7	Разработка модели машинного обучения	С	26															
8	Создание модуля обоснования принятия решения	С	26															
9	Согласование выполненной работы с научным руководителем	Р	1															
10	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	С	7															
11	Подведение итогов, оформление работы	С	5															
		Р	1															

Руководитель (Р)

Студент (С)

В итоге руководитель потратит на проект около 5 дней, студент – 121 день.

6.3.2 Бюджет научного исследования

Предлагается сравнить несколько исполнений. Исполнение 1 и 3 будет рассчитываться при условии, что оборудование для расчетов будет закупаться самостоятельно. Исполнение 2 рассчитано при помощи облачного сервиса, что значительно снизит траты на амортизацию оборудования (из-за уменьшения требований к рабочей машине), однако время разработки не изменится, так как все проводимые исследования останутся теми же.

Стоимость аренды облака – 38370 рублей в месяц. Необходим месяц непрерывных вычислений.

В рамках исполнения 3 рассмотрим вариант с использованием готовых фреймворков, которые значительно снизят время разработки ценой ухудшения качества работы системы и её меньшей специализации на решаемой задаче. Затраченные часы для руководителя не изменятся. Затраченные часы для студента – 102 рабочих дня.

Группировку затрат по статьям представлены в таблице 6.9.

Таблица 6.9 – Статьи затрат на НТИ

Статьи				
Варианты исполнения	Специальное оборудование, руб.	Основная заработная плата, руб.	Отчисления на социальные нужды, руб.	Итого
1	99999	132 371.79	39 711.54	272082,33
2	38370	132 371.79	39 711.54	210453,33
3	99999	113 144.93	33 943.48	247087,41

Таблица 6.10 – Расчет затрат на «Спецоборудование для научных работ».

№	Наименование оборудование	Кол-во	Цена, руб.	Стоимость, руб.
Вариант исполнения 1 и 3				
1	Персональный компьютер	1	99999	99999
2	Среда разработки JetBrains PyCharm	1	-	-

Итого:				99999
Вариант исполнения 2				
1	Аренда облака	1	38370	38370
2	Среда разработки JetBrains PyCharm	1	-	-
Итого:				38370

Должность руководителя – профессора, д.т.н. – 37300 рублей в месяц. Должность инженера – студента – 16242 рублей в месяц (по МРОТ).

Месячная заработная плата рассчитывается по формуле:

$$Зосн = Здн * Тр,$$

где Зосн – среднедневная зарплата, руб.;

Тр – продолжительность работ, выполняемых работником, раб. дни. Среднемесячная зарплата рассчитывается по формуле:

$$Здн = \frac{Зм * kp * М}{F_д},$$

где Зм – месячный оклад работника, руб.;

М – количество месяцев работы без отпуска в течении года: (10.4);

$F_д$ – действительный годовой фонд рабочего времени персонала (представлен в таблице 6.11).

Таблица 6.11 – Баланс рабочего времени участников разработки

Показатели рабочего времени	Руководитель	Инженер
Календарное число дней	365	365
Количество нерабочих дней – выходные дни – праздничные дни	 52 11	 82 14
Потери рабочего времени – отпуск – невыходы по болезни	 56 -	 24 -
	254	217

Среднемесячная зарплата руководителя составляет:

$$Здн = \frac{Зм * kp * М}{F_д} = \frac{37300 * 1.3 * 10.4}{254} = 1985.41$$

Среднемесячная зарплата студента составляет:

$$Здн = \frac{Зм * kp * М}{F_д} = \frac{16242 * 1.3 * 10.4}{217} = 1011.94$$

Расчет основной зарплаты представлен в таблице 6.12.

Таблица 6.12 – Расчет основной заработной платы

Исполнители	$З_{ок}$, руб.	k_p	$З_м$, руб.	$З_{дн}$, руб.	Тр, раб. дн.	$З_{осн}$, руб.	Отчисления в социальные внебюджетные фонды 30 %
Варианты исполнения 1 и 2							
Руководитель	37300	1.3	48490	1985.41	5	9 927.05	2 978.12
Инженер	16242	1.3	21114.6	1011.94	121	122 444.74	36 733.42
Итого по статье $З_{осн}$						132 371.79	39 711.54
Вариант исполнения 3							
Руководитель	37300	1.3	48490	1985.41	5	9 927.05	2 978.12
Инженер	16242	1.3	21114.6	1011.94	102	103 217.88	30965.36
Итого по статье $З_{осн}$						113 144.93	33943.48

6.4. Определение сравнительной эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат трех или более вариантов исполнения научного исследования. Для этого наибольший интегральный показатель реализации технической задачи принимается за базу расчета как знаменатель, с которым соотносится финансовые значения по всем вариантам исполнения.

Интегральный финансовый показатель разработки определяется как:

$$I_{\text{ф}}^{p} = \frac{\Phi_{pi}}{\Phi_{max}},$$

где $I_{\text{ф}}^{\text{p}}$- интегральный финансовый показатель разработки;

Φ_{pi} – стоимость i-го варианта исполнения;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

$$I_{\text{ф1}}^{p} = \frac{\Phi_{p1}}{\Phi_{max}} = \frac{272082{,}33}{272082{,}33} = 1$$

$$I_{\text{ф2}}^{p} = \frac{\Phi_{p2}}{\Phi_{max}} = \frac{210453{,}33}{272082{,}33} = 0{,}77$$

$$I_{\text{ф3}}^{p} = \frac{\Phi_{p3}}{\Phi_{max}} = \frac{247087{,}41}{272082{,}33} = 0{,}91$$

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в разах (значение больше единицы), либо соответствующее численное удешевление стоимости разработки в разах (значение меньше единицы, но больше нуля).

Определенная величина показывает, что для реализации был выбран относительно недорогой метод. Необходимо понять, изменение потребительских характеристик различных методов, чтобы окончательно понять, насколько выбранное исполнение эффективно.

Можно определить следующим образом:

$$I_{pi} = \sum_{i=1}^{n} a_i * b_i^p,$$

где Ipi – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки;

ai – весовой коэффициент i-го варианта исполнения разработки;

$bi\,p$ – бальная оценка i-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания;

Расчет приведен в таблице 6.15.

Таблица 6.15– Сравнение характеристик вариантов исполнения проекта

Критерии \ ПО	Вес	Исп. 1	Исп. 2	Исп. 3
Удобство использования	0,19	5	4	3
Масштабируемость	0,15	4	3	3
Требуемые ресурсы	0,15	4	2	4
Функциональность	0,15	4	4	4
Удобство обслуживания	0,07	5	3	3
Срок разработки	0,07	4	2	5
Надёжность	0,19	5	4	3
Итого (сумма)	1	4,33	3,22	3,35

Сравнение интегрального показателя эффективности текущего проекта и аналогов позволит определить сравнительную эффективность проекта.

Таблица 6.16 – Сравнительные показатели эффективности разработки

№	Показатели	Исп. 1	Исп. 2	Исп. 3
1	Интегральныйфинансовый показатель разработки	1	0,77	0,91
2	Интегральный показатель ресурсоэффективности разработки	4,33	3,22	3,35
3	Интегральныйпоказатель эффективности	4,33	4,18	3,68
4	Сравнительная эффективность аналогов и разработки		1,04	1,18

Сравнение значений интегральных показателей эффективности позволяет судить о приемлемости существующего варианта решения поставленной в магистерской диссертации технической задачи с позиции финансовой и ресурсной эффективности. По результатам проведенных расчетов первый вариант оказался наиболее эффективным в сравнении с другими исполнениями на 18%. Качество работы сопоставимо со вторым

исполнением. В свою очередь третье исполнение несмотря на то, что оно дешевле, проигрывает в качестве почти по всем пунктам.

Сравнение конкурентных технических решений показало, что разрабатываемая система имеет лучшие качества.

Проведена оценка готовности проекта к коммерциализации, которая показала, что перспективность разработки средняя.

Рассчитан интегральный финансовый показатель в ходе оценки бюджета затрат трех вариантов исполнения. Интегральный показатель реализации разработки равен 4,33. Рассчитан интегральный показатель ресурсоэффективности для трех вариантов исполнения: 3,22, 3,35. Расчет интегрального показатель эффективности для разработки и аналога позволил рассчитать сравнительную эффективность разработки, которая говорит о приемлемости существующего варианта решения поставленной в магистерской диссертации технической задачи с позиции финансовой и ресурсной эффективности.

7. Социальная ответственность

Введение

Результатом выполнения дипломной работы является система для семантического сравнения двух текстов на основе нейросетевого алгоритма с применением архитектуры BERT. Разработанные в результате работы алгоритмы могут быть применены отдельно или как часть большей системы. Использование данной системы позволяет единственному человеку при помощи ЭВМ обрабатывать большие объемы информации без значительных временных затрат при помощи предварительно обученных моделей. Потенциальными пользователями системы могут быть поисковые системы, системы для обработки текстов и рекомендаций. География использования системы не ограничена.

В данном разделе рассматриваются опасные и вредные факторы, оказывающие негативное влияние на разработчика системы, пользователя, окружающей среды и общества в целом. Так же рассматриваются правовые вопросы и организационные мероприятия при чрезвычайных ситуациях.

Система во время разработки и работы не должна наносить вред персоналу, пользователям, окружающей среде и как-либо нарушать действующее законодательство.

Разработка программных средств происходит при помощи ПЭВМ (персональная электронная вычислительная машина) в офисном помещении с использованием сети Интернет.

7.1.Производственная безопасность

7.1.1. Вредные факторы

7.1.1.1. Недостаточная освещенность рабочей зоны

Плохое естественное и искусственное освещение рабочего места оказывает влияние на физическое и психологическое состояние пользователя,

что неблагоприятно сказывается на его работе. Не надлежащее качество освещения может привести к ухудшению зрения.

Согласно СП 52.13330.2016 [25] при работах III зрительного разряда и подразряда г (работы высокой точности) освещенность при системе общего освещения должна быть не ниже 200 Лк.

Расчет общего равномерного искусственного освещения горизонтальной рабочей поверхности выполняется методом коэффициента использования светового потока, учитывающим световой поток, отраженный от потолка и стен. Длина помещения $A = 6$м, ширина В $= 3$м, высота $H = 3$м. Высота рабочей поверхности над полом $h_p = 0.8$м.

Площадь помещения:

$$S = A * B = 6 * 3 = 18\text{м}^2$$

Коэффициент отражения стен, оклеенных светлыми обоями с окнами, без штор $p_c = 30\%$, потолка светлой поверхности $p_{\text{п}} = 50\%$. Коэффициент запаса, учитывающий загрязнение светильника, для помещений с малым выделением пыли равен $K_{\text{з}} = 1{,}5$. Коэффициент неравномерности для люминесцентных ламп $Z = 1{,}1$.

Согласно нормативным документам, принятым в предприятии, помещение освещено тремя светильниками типа ОДОР-2-40 с двумя лампами ЛД-40 со световым потоком $\Phi_{\text{п}} = 2300$ Лм. По паспорту длина светильника $\text{А}_{\text{св}} = 1227$мм, ширина $B_{\text{св}} = 265$мм. Мощность лампы – 40 Вт.

Интегральным критерием оптимальности расположения светильников является величина λ, которая для люминесцентных светильников с защитной решеткой лежит в диапазоне 1,1-1,3. Примем $\lambda = 1{,}2$.

Расстояние светильников от перекрытия (свес): $h_c = 0{,}4$ м.

Высота светильника над рабочей поверхностью определяется по формуле:

$$h = H - h_p - h_c = 3 - 0{,}8 - 0{,}4 = 1{,}8\text{м}$$

Индекс помещения определяется по формуле:

$$i = \frac{A * B}{h * (A + B)} = \frac{6 * 3}{1{,}8 * (6 + 3)} = 1{,}11$$

Из методической таблицы коэффициент использования светового потока, показывающий какая часть светового потока ламп попадает на рабочую поверхность, для светильников типа ОДОР с люминесцентными лампами при $p_c = 30\%$, $p_{\text{п}} = 50\%$ и индексе помещения i = 1,11 равен $\eta = 39\% = 0{,}39$.

План помещения и размещения светильников с люминесцентными лампами представлен на рисунке 7.1.

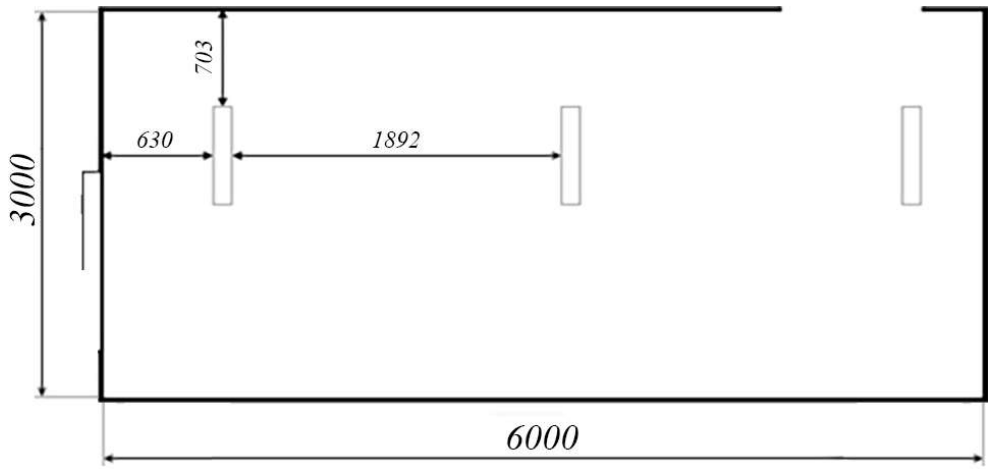

Рисунок 7.1 – схема расположения светильников с люминесцентными лампами в производственном помещении

Общее число светильников: $N_{\text{св}} = 3$. Соответственно количество люминесцентных ламп: $N_{\text{лл}} = 6$.

Расчет светового потока группы люминесцентных ламп светильника определяется по формуле:

$$\Phi_{\text{рас}} = \frac{E * A * B * K_{\text{З}} * Z}{N_{\text{лл}} * \eta} = \frac{200 * 6 * 3 * 1{,}5 * 1{,}1}{6 * 0{,}39} = 2538$$

Делаем проверку выполнения условия:

$$-10\% \leq \frac{\Phi_{\text{п}} - \Phi_{\text{рас}}}{\Phi_{\text{лд}}} * 100\% \leq 20\%$$

$$\frac{\Phi_{\text{п}} - \Phi_{\text{рас}}}{\Phi_{\text{лд}}} * 100\% = \frac{2300 - 2538}{2300} * 100\% = -10\%$$

Таким образом: $-10\% \leq -10\% \leq 20\%$, необходимый световой поток светильника не выходит за пределы требуемого диапазона, хотя и находится на грани.

Мощность осветительной установки рассчитывается по формуле:

$$P_{\text{уст}} = N * P_{\text{л}} = 6 * 40 = 240\text{Вт}.$$

Рассчитаем удельную мощность осветительной установки по формуле:

$$P_{\text{уд}} = \frac{\text{N} * \text{P}_{\text{л}}}{S} = \frac{6 * 40}{18} = 13{,}3 \ \frac{\text{Вт}}{\text{м}^2}$$

Далее перечислены общие требования и рекомендации к организации освещения на рабочем месте.

- рабочие места следует размещать таким образом, чтобы естественный свет падал преимущественно слева, а дисплеи мониторов были ориентированы боковой стороной к световым проемам;
- система общего равномерного освещения должна регулировать искусственное освещение в помещениях для эксплуатации ПЭВМ.

Вышеперечисленные меры полностью соблюдаются, что позволяет сохранить зрение и избежать пагубного воздействия на глаза во время разработки и эксплуатации результатов выпускной квалификационной работы.

7.1.1.2. Отклонение показателей микроклимата

Комфортные условия для работы создаются оптимальным сочетанием температуры, относительной влажности и скорости движения воздуха. На рабочих местах пользователей ПЭВМ должны обеспечиваться оптимальные параметры микроклимата в соответствии с СанПиН 1.2.3685-21 [26]

Согласно этому документу, должны быть соблюдены требования, описанные в таблицах 7.1 и 7.2.

Таблица 7.1 – Оптимальные нормы микроклимата

Период года	Температура воздуха, С°	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	19-23	40-60	0.1
Теплый	23-25		0.2

Таблица 7.2 - Допустимые нормы микроклимата

Период года	Температура воздуха, С°		Относительная влажность воздуха, %	Скорость движения воздуха, м/с
	Нижняя допустимая граница	Верхняя допустимая граница		
Холодный	15	24	20-80	<0.5
Теплый	22	28	20-80	<0.5

Для поддержания оптимальных значений микроклимата используется системы отопления и кондиционирования воздуха, тепловая изоляция нагретых поверхностей оборудования. При исследовании микроклимата было выявлено, что в кабинетах, где выполнялась разработка, параметры микроклимата соответствуют требованиям СанПиН.

7.1.1.3. Превышение уровня шума

Шум является одним из распространенных в производстве вредных факторов. Его создают работающее оборудование, преобразователи напряжения, работающие осветительные приборы дневного света и другие источники шума. Шум может стать причиной снижения работоспособности и

повышенной утомляемости. Значительные превышения уровня шума на рабочем месте вызывают необратимые изменения в органах слуха человека, также оказывают неблагоприятное влияние на весь организм человека через нервную систему. В результате ослабляется внимание, ухудшается память, снижается реакция, что вызывает увеличение числа ошибок при работе.

Требования к допустимому уровню шума были описаны в ГОСТ 12.1.003-83 [27] и СанПиН 1.2.3685-21 [26]. Согласно данному документу, максимально допустимый уровень шума составляет не более 65 дБа, если осуществляется умственная деятельность, которая требует постоянной концентрации.

В процессе выполнения проекта на улице около производственного помещения проводились непродолжительные ремонтные работы, во время которых шумовое загрязнение значительно превышало норму. В рамках улучшения условий труда офис был временно изменен на аналогичное помещение с другой стороны здания, в котором уровень шума не превышал допустимый уровень.

Все источники громкого шума (принтеры, сервера и другое оборудование) были изолированы в отдельном помещении.

При значениях выше допустимого уровня необходимо предусмотреть средства индивидуальной защиты (СИЗ) и средства коллективной защиты (СКЗ) от шума.

Средства коллективной защиты:

1. устранение причин шума или существенное его ослабление в источнике образования;

2. изоляция источников шума от окружающей среды (применение глушителей, экранов, звукопоглощающих строительных материалов, например

любой пористый материал – шамотный кирпич, микропористая резина, поролон и др.);

3. применение средств, снижающих шум и вибрацию на пути их распространения;

Средства индивидуальной защиты:

1. применение спецодежды и защитных средств органов слуха: наушники, беруши, антифоны.

7.1.1.4. Повышенный уровень электромагнитного излучения, ПДУ, СКЗ, СИЗ

Источником электромагнитных излучений в нашем случае являются дисплеи ПЭВМ. Монитор компьютера включает в себя излучения рентгеновской, ультрафиолетовой и инфракрасной области, а также широкий диапазон электромагнитных волн других частот. Согласно СанПиН 2.2.4.3359-16 [28] напряженность электромагнитного поля по электрической составляющей на расстоянии 50 см вокруг ВДТ не должна превышать 25В/м в диапазоне от 5Гц до 2кГц, 2,5В/м в диапазоне от 2 до 400кГц [28]. Плотность магнитного потока не должна превышать в диапазоне от 5 Гц до 2 кГц 250нТл, и 25нТл в диапазоне от 2 до 400кГц. Поверхностный электростатический потенциал не должен превышать 500В [28]. В ходе работы использовалась ПЭВМ типа Acer VN7-791 со следующими характеристиками: напряженность электромагнитного поля 2,5В/м; поверхностный потенциал составляет 450 В (основы противопожарной защиты предприятий ГОСТ 12.1.004 [30] и ГОСТ 12.1.010 – 76 [31]).

При длительном постоянном воздействии электромагнитного поля (ЭМП) радиочастотного диапазона при работе на ПЭВМ у человеческого организма сердечно-сосудистые, респираторные и нервные расстройства, головные боли, усталость, ухудшение состояния здоровья, гипотония, изменения сердечной мышцы проводимости. Тепловой эффект ЭМП

характеризуется увеличением температуры тела, локальным селективным нагревом тканей, органов, клеток за счет перехода ЭМП на теплую энергию.

Предельно допустимые уровни (ПДУ) облучения (по ГОСТ 54 30013-83) [32]:

а) до 10 мкВт. /см2, время работы (8 часов);

б) от 10 до 100 мкВт/см2, время работы не более 2 часов;

в) от 100 до 1000 мкВт/см2, время работы не более 20 мин. при условии пользования защитными очками;

г) для населения в целом ППМ не должен превышать 1 мкВт/см2.

Защита человека от опасного воздействия электромагнитного излучения осуществляется следующими способами:

СКЗ

1. защита временем;
2. защита расстоянием;
3. снижение интенсивности излучения непосредственно в самом источнике излучения;
4. заземление экрана вокруг источника;
5. защита рабочего места от излучения;

СИЗ

1. Очки и специальная одежда, выполненная из металлизированной ткани (кольчуга). При этом следует отметить, что использование СИЗ возможно при кратковременных работах и является мерой аварийного характера. Ежедневная защита обслуживающего персонала должна обеспечиваться другими средствами.
2. Вместо обычных стекол используют стекла, покрытые тонким слоем золота или диоксида олова (SnO_2).

7.1.2.1. Электроопасность; класс электроопасности помещения, безопасные номиналы I, U, $R_{заземления}$, СКЗ, СИЗ

Поражение электрическим током

К опасным факторам можно отнести наличие в помещении большого количества аппаратуры, использующей однофазный электрический ток напряжением 220 В и частотой 50 Гц. По опасности электропоражения комната относится к помещениям без повышенной опасности, так как отсутствует повышенная влажность, высокая температура, токопроводящая пыль и возможность одновременного соприкосновения токоведущих элементов с заземленными металлическими корпусами оборудования.

Лаборатория относится к помещению без повышенной опасности поражения электрическим током. Безопасными номиналами являются: I < 0,1 А; U < (2-36) В; Rзазем < 4 Ом.

Для защиты от поражения электрическим током используют СИЗ и СКЗ.

Средства коллективной защиты:

- защитное заземление, зануление;
- малое напряжение;
- электрическое разделение сетей;
- защитное отключение;
- изоляция токоведущих частей;
- оградительные устройства.

Использование щитов, барьеров, клеток, ширм, а также заземляющих и шунтирующих штанг, специальных знаков и плакатов.

Средства индивидуальной защиты:

Использование диэлектрических перчаток, изолирующих клещей и штанг, слесарных инструментов с изолированными рукоятками, указатели величины напряжения, калоши, боты, подставки и коврики.

7.1.2.2. Пожароопасность, категория пожароопасности помещения, марки огнетушителей, их назначение и ограничение применения; Приведена схема эвакуации.

По взрывопожарной и пожарной опасности помещения подразделяются на категории А, Б, В1-В4, Г и Д.

Согласно НПБ 105-03 лаборатория относится к категории В – горючие и трудно горючие жидкости, твердые горючие и трудно горючие вещества и материалы, вещества и материалы, способные при взаимодействии с водой, кислородом воздуха или друг с другом только гореть, при условии, что помещения, в которых находится, не относятся к категории наиболее опасных А или Б.

По степени огнестойкости данное помещение относится к 1-й степени огнестойкости по СНиП 2.01.02-85 [33] (выполнено из кирпича, которое относится к трудносгораемым материалам).

Возникновение пожара при работе с электронной аппаратурой может быть по причинам как электрического, так и неэлектрического характера.

Причины возникновения пожара неэлектрического характера:

а) халатное неосторожное обращение с огнем (курение, оставленные без присмотра нагревательные приборы, использование открытого огня);

Причины возникновения пожара электрического характера: короткое замыкание, перегрузки по току, искрение и электрические дуги, статическое электричество и т. п.

Для локализации или ликвидации загорания на начальной стадии используются первичные средства пожаротушения. Первичные средства пожаротушения обычно применяют до прибытия пожарной команды.

Огнетушители водо-пенные (ОХВП-10) используют для тушения очагов пожара без наличия электроэнергии. Углекислотные (ОУ-2) и порошковые огнетушители предназначены для тушения электроустановок, находящихся под напряжением до 1000В. Для тушения токоведущих частей и

электроустановок применяется переносной порошковый огнетушитель, например ОП-5.

В общественных зданиях и сооружениях на каждом этаже должно размещаться не менее двух переносных огнетушителей. Огнетушители следует располагать на видных местах вблизи от выходов из помещений на высоте не более 1,35 м. Размещение первичных средств пожаротушения в коридорах, переходах не должно препятствовать безопасной эвакуации людей.

Для предупреждения пожара и взрыва необходимо предусмотреть:

1. специальные изолированные помещения для хранения и разлива легковоспламеняющихся жидкостей (ЛВЖ), оборудованные приточно-вытяжной вентиляцией во взрывобезопасном исполнении - соответствии с ГОСТ 12.4.021-75 [34] и СНиП 2.04.05-86 [35];

2. специальные помещения (для хранения в таре пылеобразной канифоли), изолированные от нагревательных приборов и нагретых частей оборудования;

3. первичные средства пожаротушения на производственных участках (передвижные углекислые огнетушители, пенные огнетушители ТУ 22-4720-80, ящики с песком, войлок, кошма или асбестовое полотно);

4. автоматические сигнализаторы (типа СВК-3 М 1) для сигнализации о присутствии в воздухе помещений предвзрывных концентраций горючих паров растворителей и их смесей.

Лаборатория полностью соответствует требованиям пожарной безопасности, а именно, наличие охранно-пожарной сигнализации, плана эвакуации, изображенного на рисунке 7.2, порошковых огнетушителей с поверенным клеймом, табличек с указанием направления к запасному (эвакуационному) выходу.

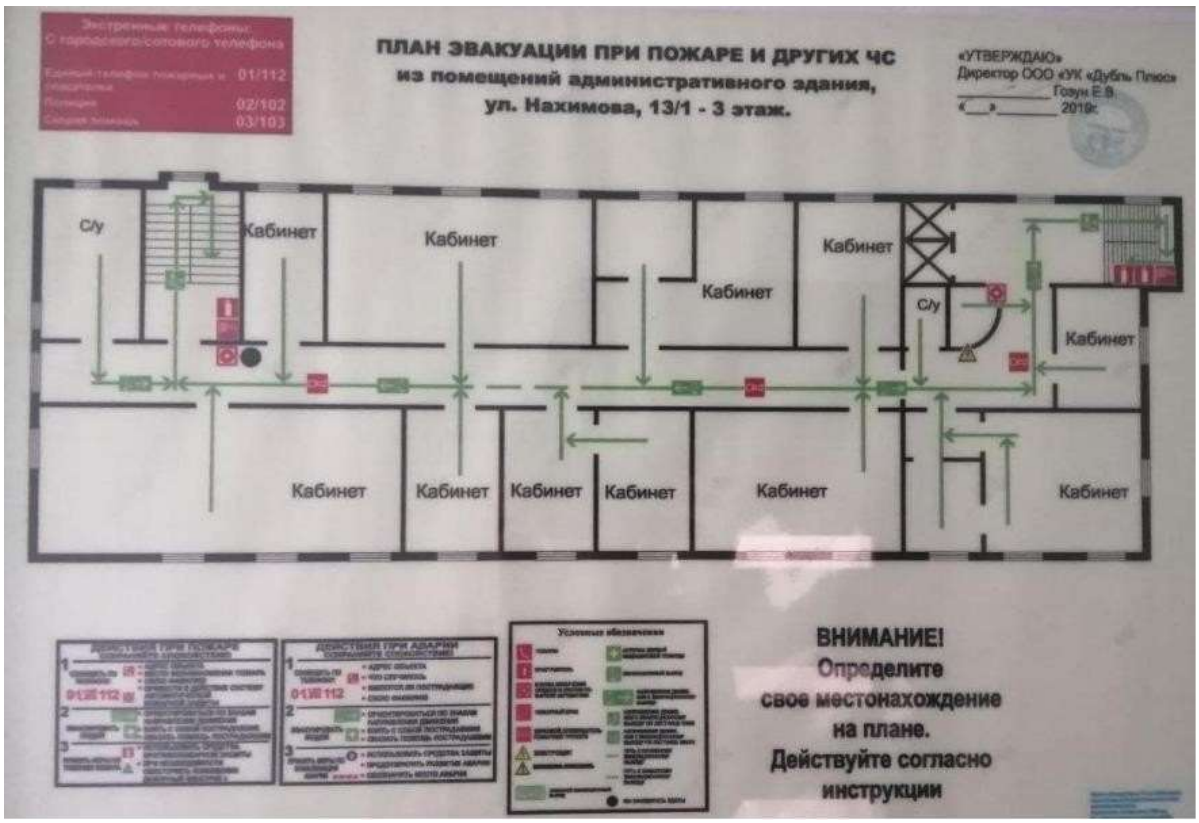

Рисунок 7.2 – План эвакуации

7.2. Экологическая безопасность

Этот раздел посвящен вопросам наличия и утилизации промышленных отходов: бумаги-черновиков, вторцвета и чермета, пластмассы, перегоревших люминесцентных ламп, а также вышедшей из строя оргтехники.

Бумага-черновики, как правило, подлежат переработке. Большинство производств применяют программы, в которых отходы бумаги собираются и перерабатываются, что снижает негативное воздействие на окружающую среду.

Вторцвет и чермет, будучи важными источниками вторичного сырья, могут быть повторно использованы для создания новых продуктов. Обычно отходы этих видов отправляются на специализированные заводы, где они сортируются и перерабатываются.

Пластмассовые отходы являются значительной проблемой для окружающей среды из-за их долгого периода разложения. Но их также можно эффективно утилизировать путем механической или химической переработки.

Процессы могут включать дробление, плавление, формовку и другие виды обработки для создания новых пластмассовых изделий.

Перегоревшие люминесцентные лампы и оргтехника содержат ряд потенциально опасных веществ и требуют особого обращения. Эти отходы собираются и отправляются на специализированные предприятия, где проводится безопасная утилизация и нейтрализация вредных веществ. Люминесцентные лампы утилизируют следующим образом. Не работающие лампы немедленно после удаления из светильника должны быть упакованы в картонную коробку, бумагу или тонкий мягкий картон, предохраняющий лампы от взаимного соприкосновения и случайного механического повреждения. После накопления ламп объемом в 1 транспортную единицу их сдают на переработку на соответствующее предприятие. Недопустимо выбрасывать отработанные энергосберегающие лампы вместе с обычным мусором, превращая его в ртутьсодержащие отходы, которые загрязняют ртутными парами

Таким образом, учёт и обеспечение эффективной утилизации промышленных отходов играют ключевую роль в поддержании экологической безопасности.

7.3. Безопасность в чрезвычайных ситуациях

Природная чрезвычайная ситуация – обстановка на определенной территории или акватории, сложившейся в результате возникновения источника природной чрезвычайной ситуации, который может повлечь или повлек за собой человеческие жертвы, ущерб здоровью людей и (или) окружающей природной среде, значительные материальные потери и нарушение условий жизнедеятельности людей.

Производство находится в городе Томске с континентально-циклоническим климатом. Природные явления (землетрясения, наводнения, засухи, ураганы и т. д.), в данном городе отсутствуют.

Возможными ЧС на объекте в данном случае, могут быть сильные морозы и диверсия.

Для Сибири в зимнее время года характерны морозы. Достижение критически низких температур приводит к авариям систем тепло- и водоснабжения, сантехнических коммуникаций и электроснабжения, приостановке работы. В этом случае при подготовке к зиме следует предусмотреть а) газобаллонные калориферы (запасные обогреватели), б) дизель или бензоэлектрогенераторы; в) запасы питьевой и технической воды на складе (не менее 30 л на 1 человека); г) теплый транспорт для доставки работников на работу и с работы домой в случае отказа муниципального транспорта. Их количества и мощности должно хватать для того, чтобы работа на производстве не прекратилась.

Для предупреждения вероятности осуществления диверсии предприятие необходимо оборудовать системой видеонаблюдения, круглосуточной охраной, пропускной системой, надежной системой связи, а также исключения распространения информации о системе охраны объекта, расположении помещений и оборудования в помещениях, системах охраны, сигнализаторах, их местах установки и количестве. Должностные лица раз в полгода проводят тренировки по отработке действий на случай экстренной эвакуации.

Перечень НТД

1. СП 52.13330.2016. Естественное и искусственное освещение.
2. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.
3. ГОСТ 12.1.003-83. Система стандартов безопасности труда. Шум. Общие требования безопасности.
4. СанПиН 2.2.4.3359-16. Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.
5. СП 2.4.3648-20. Санитарно-эпидемиологические требования к организации воспитания и обучения, отдыха и оздоровления детей и молодежи.
6. ГОСТ 12.1.004. Система стандартов безопасности труда. Пожарная безопасность.
7. ГОСТ 12.1.010 – 76. Система стандартов безопасности труда. Взрывобезопасность.
8. ГОСТ 54 30013-83. Электромагнитные излучения СВЧ. Предельно допустимые уровни облучения. Требования безопасности.
9. СНиП 2.01.02-85. Противопожарные нормы.
10. ГОСТ 12.4.021-75. Система стандартов безопасности труда. Системы вентиляционные.
11. СНиП 2.04.05-86. Отопление, вентиляция и кондиционирование.
12. ГОСТ 12.4.154-85. ССБТ. Устройства, экранирующие для защиты от электрических полей промышленной частоты.
13. ГН 2.2.5.1313-03. Предельно допустимые концентрации (ПДК) вредных веществ в воздухе рабочей зоны.
14. СанПиН 2.2.4/2.1.8.055-96. Электромагнитные излучения радиочастотного диапазона (ЭМИ РЧ).
15. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений.

16. СН 2.2.4/2.1.8.562-96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки.
17. ГОСТ 12.4.123-83. Средства коллективной защиты от инфракрасных излучений. Общие технические требования.
18. ГОСТ Р 12.1.019-2009. Электробезопасность. Общие требования и номенклатура видов защиты.
19. ГОСТ 12.1.030-81. Электробезопасность. Защитное заземление. Зануление.
20. ГОСТ 12.1.004-91. Пожарная безопасность. Общие требования.
21. ГОСТ 12.2.037-78. Техника пожарная. Требования безопасности.
22. СанПиН 2.1.6.1032-01. Гигиенические требования к качеству атмосферного воздуха.
23. ГОСТ 30775-2001. Ресурсосбережение. Обращение с отходами. Классификация, идентификация и кодирование отходов.
24. СНиП 21-01-97. Противопожарные нормы.
25. ГОСТ 12.4.154. Система стандартов безопасности труда. Устройства, экранирующие для защиты от электрических полей промышленной частоты. Общие технические требования, основные параметры и размеры.

Заключение

В процессе выполнения этой работы проводилась проработка исследовательской работы, связанной с семантическим сравнением текстов при помощи методов машинного обучения. Обзор был сосредоточен на изучении различных методов и подходов к семантическому сравнению текстов, с акцентом на архитектуру BERT, представляющую собой одну из наиболее эффективных и адаптивных моделей для этой цели. Разнообразные варианты использования BERT для семантического сравнения текстов были изучены, включая косинусное и евклидово расстояния, а также применение обученной нейронной сети к векторам.

В рамках данного исследования реализовывались различные алгоритмы машинного обучения для семантического сравнения текстов, основанные на архитектуре BERT. Эксперименты были направлены на оптимизацию этих алгоритмов и улучшение их производительности. При этом были разработаны три модели. Первая модель, основанная на RuBERT, использовала обученную нейронную сеть для предсказания на основе косинусных расстояний между векторами внутренних представлений двух текстов, и показала среднеквадратичную ошибку на тестовой выборке равную 0.0656. Вторая модель, базирующаяся на модели BERT paraphrase-multilingual-MiniLM-L12-v2, использовала нейронную сеть, обученную предсказывать на основе векторных представлений, и показала результат среднеквадратичной ошибки равный 0.05. Третья модель, основанная на нейронной сети, обученной определять схожесть текстов по евклидовым расстояниям, показала результат среднеквадратичной ошибки равный 0.019.

Модель, обученная на основе евклидовых расстояний, оказалась наиболее точной в предсказании семантической близости текстов, превосходя модель, обученную на векторах, и в особенности первоначальную модель, основанную на RuBERT и использующую сравнение косинусных расстояний, с среднеквадратичной ошибкой 0.0656.

Так, результаты экспериментов подтверждают, что предложенный подход обеспечивает высокую точность семантического сравнения текстов и может быть успешно применен в задачах обработки естественного языка. Результаты данной работы подчеркивают актуальность и важность применения методов машинного обучения для семантического сравнения текстов, и в то же время иллюстрируют потенциал архитектуры BERT как одного из наиболее эффективных инструментов в этой области. В будущем эти результаты могут быть использованы для создания более продвинутых систем обработки естественного языка, способных к эффективному анализу и сравнению текстов на семантическом уровне.

Произведен обзор и оценка проведенных научных исследований с точки зрения финансово-экономической перспективы, с учетом полных денежных затрат, выделенных на реализацию проекта.

Исследовались негативные и опасные факторы, способные оказать негативное воздействие на здоровье. Проанализированы вопросы, относящиеся к экологической безопасности и охране труда на рабочем месте. Выполнен анализ вредных и опасных факторов при взаимодействии с системой.

Список источников

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All you Need [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1706.03762 (дата обращения 17.05.2023).
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1810.04805 (дата обращения 17.05.2023).
3. Nils Reimers, Iryna Gurevych (2018). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1908.10084 (дата обращения 17.05.2023).
4. Daniel Cer, Yinfei Yang, Sheng-yi Kong (2018). Universal Sentence Encoder [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1908.10084 (дата обращения 17.05.2023).
5. Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, Lucia Specia (2018). SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs//1708.00055 (дата обращения 17.05.2023).

   Paraphrase Database (PPDB) [Электронный ресурс]: http://paraphrase.org [сайт]. Режим доступа: http://paraphrase.org (дата обращения 17.05.2023).
6. SICK (Sentences Involving Compositional Knowledge) Dataset [Электронный ресурс]: https://zenodo.org [сайт]. Режим доступа: https://zenodo.org/record/2787612 (дата обращения 17.05.2023).
7. MSR (Microsoft Research) Paraphrase Corpus [Электронный ресурс]: https://www.microsoft.com [сайт]. Режим доступа: https://www.microsoft.com/en-us/download/details.aspx?id=52398 (дата обращения 17.05.2023).

8. MAP метрика (Mean Average Precision) [Электронный ресурс]: https://en.wikipedia.org [сайт]. Режим доступа: https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision (дата обращения 17.05.2023).
9. RuBERT (Russian BERT) [Электронный ресурс]: https://github.com/deepmipt/DeepPavlov [сайт]. Режим доступа: https://github.com/deepmipt/DeepPavlov (дата обращения 17.05.2023).
10. distilbert-base-multilingual-cased [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/distilbert-base-multilingual-cased (дата обращения 17.05.2023).
11. sentence-transformers/distiluse-base-multilingual-cased-v1 [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1 (дата обращения 17.05.2023).
12. inkoziev/sbert_synonymy [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/inkoziev/sbert_synonymy (дата обращения 17.05.2023).
13. sentence-transformers/stsb-xlm-r-multilingual [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual (дата обращения 17.05.2023).
14. paraphrase-multilingual-MiniLM-L12-v2 [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/paraphrase-multilingual-MiniLM-L12-v2 (дата обращения 17.05.2023).
15. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1908.10084 (дата обращения 17.05.2023)

16. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1909.11942 (дата обращения 17.05.2023).
17. Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1909.00161 (дата обращения 17.05.2023).
18. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/1908.10084 (дата обращения 17.05.2023).
19. Li, Y., Li, W., & Chao, H. (2020). A Novel Siamese Bert-Based Network for Semantic Sentence Matching [Электронный ресурс]: https://arxiv.org [сайт]. Режим доступа: https://arxiv.org/abs/2004.01992 (дата обращения 17.05.2023).
20. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning [Электронный ресурс]: https://www.nature.com [сайт]. Режим доступа: https://www.nature.com/articles/nature14539 (дата обращения 17.05.2023).
21. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.) [Электронный ресурс]: https://web.stanford.edu/~jurafsky/slp3/ [сайт]. Режим доступа: https://web.stanford.edu/~jurafsky/slp3/ (дата обращения 17.05.2023).
22. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing [Электронный ресурс]: http://nlp.stanford.edu/fsnlp/ [сайт]. Режим доступа: http://nlp.stanford.edu/fsnlp/ (дата обращения 17.05.2023).
23. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit [Электронный

ресурс]: http://www.nltk.org/book/ [сайт]. Режим доступа: http://www.nltk.org/book/ (дата обращения 17.05.2023).

24. Eisenstein, J. (2018). Introduction to Natural Language Processing [Электронный ресурс]: http://comp.social.gatech.edu/papers/ [сайт]. Режим доступа: http://comp.social.gatech.edu/papers/ (дата обращения 17.05.2023).
25. СП 52.13330.2016. Естественное и искусственное освещение.
26. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.
27. ГОСТ 12.1.003-83. Система стандартов безопасности труда. Шум. Общие требования безопасности.
28. СанПиН 2.2.4.3359-16. Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.
29. СП 2.4.3648-20. Санитарно-эпидемиологические требования к организации воспитания и обучения, отдыха и оздоровления детей и молодежи.
30. ГОСТ 12.1.004. Система стандартов безопасности труда. Пожарная безопасность.
31. ГОСТ 12.1.010 – 76. Система стандартов безопасности труда. Взрывобезопасность.
32. ГОСТ 54 30013-83. Электромагнитные излучения СВЧ. Предельно допустимые уровни облучения. Требования безопасности.
33. СНиП 2.01.02-85. Противопожарные нормы.
34. ГОСТ 12.4.021-75. Система стандартов безопасности труда. Системы вентиляционные.
35. СНиП 2.04.05-86. Отопление, вентиляция и кондиционирование.
36. ГОСТ 12.4.154-85. ССБТ. Устройства, экранирующие для защиты от электрических полей промышленной частоты.

37. ГН 2.2.5.1313-03. Предельно допустимые концентрации (ПДК) вредных веществ в воздухе рабочей зоны.
38. СанПиН 2.2.4/2.1.8.055-96. Электромагнитные излучения радиочастотного диапазона (ЭМИ РЧ).
39. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений.
40. СН 2.2.4/2.1.8.562-96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки.
41. ГОСТ 12.4.123-83. Средства коллективной защиты от инфракрасных излучений. Общие технические требования.
42. ГОСТ Р 12.1.019-2009. Электробезопасность. Общие требования и номенклатура видов защиты.
43. ГОСТ 12.1.030-81. Электробезопасность. Защитное заземление. Зануление.
44. ГОСТ 12.1.004-91. Пожарная безопасность. Общие требования.
45. ГОСТ 12.2.037-78. Техника пожарная. Требования безопасности.
46. СанПиН 2.1.6.1032-01. Гигиенические требования к качеству атмосферного воздуха.
47. ГОСТ 30775-2001. Ресурсосбережение. Обращение с отходами. Классификация, идентификация и кодирование отходов.
48. СНиП 21-01-97. Противопожарные нормы.
49. ГОСТ 12.4.154. Система стандартов безопасности труда. Устройства, экранирующие для защиты от электрических полей промышленной частоты. Общие технические требования, основные параметры и размеры.

Приложение А

Раздел 2

Text similarity comparison algorithms based on neural network algorithms

Студент:

Группа	ФИО	Подпись	Дата
8ВМ13	Цыденов Саян Баирович		

Руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ ИШИТР	Спицын Владимир Григорьевич	д.т.н.		

Консультант-лингвист отделения иностранных языков ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Сидоренко Татьяна Валерьевна	к.п.н.		

Introduction

With the advancement of information technology and the increase in the volume of available data, there has been a need for effective methods of analyzing and processing textual information. One of the key tasks in this field is semantic text comparison, which involves determining the degree of similarity between two texts based on their meaning. This task has a wide range of applications, including but not limited to information retrieval, recommendation systems, sentiment analysis, and machine translation.

The goal of this work is to develop and implement a machine learning algorithm for solving the task of semantic text comparison. The proposed approach is based on neural network methods, specifically the transformer architecture and the BERT (Bidirectional Encoder Representations from Transformers) model, which have demonstrated high effectiveness in natural language processing tasks.

Several methods of semantic text comparison will be examined in this work, including comparison based on cosine and Euclidean distances, as well as using a neural network trained on vectors. Each of these methods has its advantages and disadvantages, and the choice of a specific method depends on the task's specificity and available data.

An important part of this work is conducting experiments to evaluate the effectiveness of the proposed algorithm. Experiments will be performed on various datasets, and the results will be compared with the results obtained using other existing methods.

In this work, we aim not only to develop an efficient algorithm for semantic text comparison but also to shed light on important aspects of this complex and multifaceted task, which, we hope, will be useful for further research in this field.

1. Problem Description and Relevance of the Topic

Semantic text comparison is one of the key tasks in the field of natural language processing (NLP). It involves determining the degree of similarity between two texts based on their meaning, which presents a complex and multifaceted problem. On the one hand, texts can be structurally and lexically similar but differ in meaning. On the other hand, texts can be written with different words and styles but convey the same idea. Thus, the task of semantic text comparison requires considering both lexical and semantic aspects of the text.

The relevance of this topic is driven by the growing need for automating the processing of textual information. Semantic text comparison plays an important role in many applications such as information retrieval, recommendation systems, sentiment analysis, machine translation, and many others. For example, in search engines, semantic comparison can help improve search quality by enabling the system to better understand user queries and provide more relevant results. In recommendation systems, semantic comparison can help offer users content that more accurately matches their interests.

Despite significant advancements in NLP, semantic text comparison remains a challenging task that requires further research. In particular, existing methods may face difficulties when dealing with texts containing complex semantic structures, ambiguities, or domain-specific vocabulary. In this work, we aim to develop a new machine learning algorithm for semantic text comparison that can address these challenges and others.

2. Theoretical Overview

2.1. Review of Machine Learning Methods for Semantic Text Comparison

Semantic text comparison is a task where machine learning plays a crucial role. The machine learning methods used for this task range from classical approaches, such as bag-of-words and TF-IDF, to more complex and modern methods based on neural networks.

Bag-of-words and TF-IDF are simple yet effective methods for transforming text into a vector space, allowing the use of standard similarity metrics like cosine and Euclidean distance. However, these methods do not consider word order and semantic relationships between them, limiting their ability for semantic text comparison.

Neural networks offer a more powerful tool for semantic text comparison. They are capable of modeling complex dependencies and semantic structures in text, making them particularly useful for this task. One key advantage of neural networks is their ability to learn from large volumes of data, enabling them to extract deeper and more accurate semantic representations of text.

The transformer architecture, proposed in the paper "Attention is All You Need" (Vaswani et al., 2017), represents a significant breakthrough in natural language processing. Transformers use attention mechanisms to model dependencies between words in the text, allowing them to effectively process long sequences and consider the context of each word. This makes transformers particularly suitable for semantic text comparison tasks.

BERT (Bidirectional Encoder Representations from Transformers) is a model based on the transformer architecture, which was introduced in the work of Devlin et al. (2018). BERT employs bidirectional training on transformer representations, enabling it to better understand word context and semantics. BERT has achieved impressive results in various NLP tasks, including semantic text comparison.

Overall, machine learning methods, particularly those based on neural networks and transformer architectures like BERT, have shown great potential for

addressing the challenges of semantic text comparison by capturing semantic relationships and context in textual data.

2.2. Review of Transformer Architectures

The transformer architecture was first introduced in 2017 and has since become the foundation for many state-of-the-art models in natural language processing. The main idea behind transformers is the use of an attention mechanism that allows modeling dependencies between words in a text without considering their positional order.

Transformers consist of two main components: the encoder and the decoder. The encoder transforms the input text into a sequence of vector representations, each capturing the meaning of the corresponding word in the context of the entire text. The decoder then uses these representations to generate the output text, also taking the context into account.

The key component of transformers is the attention block, which computes the weights of interactions between all pairs of words in the text. This enables the model to consider the context of each word, regardless of its position in the text. This ability allows transformers to effectively process long sequences and capture complex semantic relationships between words.

Transformers also employ positional encoding to convey information about the order of words in the text. This allows the model to consider word order, even though the attention mechanism itself does not explicitly account for word position.

Since its introduction, the transformer architecture has served as the basis for many other models, such as BERT, GPT-4, and T5, which have demonstrated remarkable performance in various natural language processing tasks.

2.3. Review of BERT Architecture and Its Application for Semantic Text Comparison

BERT (Bidirectional Encoder Representations from Transformers) is a natural language processing model introduced in the paper by Devlin et al. (2018). BERT is based on the transformer architecture and utilizes bidirectional training to create powerful representations of text.

Unlike some previous models, such as GPT, which are trained to predict the next word in a text (known as "teacher-forcing"), BERT is trained on two tasks: predicting masked words in the text (the "fill-in-the-blank" task) and determining whether one sentence is a continuation of another. This bidirectional training enables BERT to better understand word context and semantics, making it particularly useful for semantic text comparison tasks.

BERT is trained on a large corpus of text and generates a vector representation for each word in the text. These vector representations can then be used for semantic text comparison, such as by computing cosine or Euclidean distances between vectors. BERT can also be fine-tuned for a specific task, allowing it to adapt to the specific requirements of the task and improve its performance.

The application of BERT for semantic text comparison has already demonstrated impressive results in various tasks, including sentence-level semantic equivalence, text classification, and document ranking. This makes BERT one of the most powerful and versatile tools for semantic text comparison in today's landscape.

2.3.1. BERT Architecture

The BERT (Bidirectional Encoder Representations from Transformers) architecture is a bidirectional pretrained encoder capable of effectively modeling contextual dependencies in text. The main idea is to pretrain the model on a large amount of unlabeled textual data and use the learned weights for further training on specific NLP tasks.

The advantage of BERT lies in its ability to learn contextual word representations, meaning understanding the meaning of a word based on its surrounding context in a sentence. This is achieved through the self-attention mechanism, which allows the model to consider interactions between different words in a sentence.

BERT is trained on two tasks: next sentence prediction and masked word prediction. In the first task, the model is trained to predict whether one sentence is a continuation of another. In the second task, random words in a sentence are masked or replaced with random words, and the model must predict the original word.

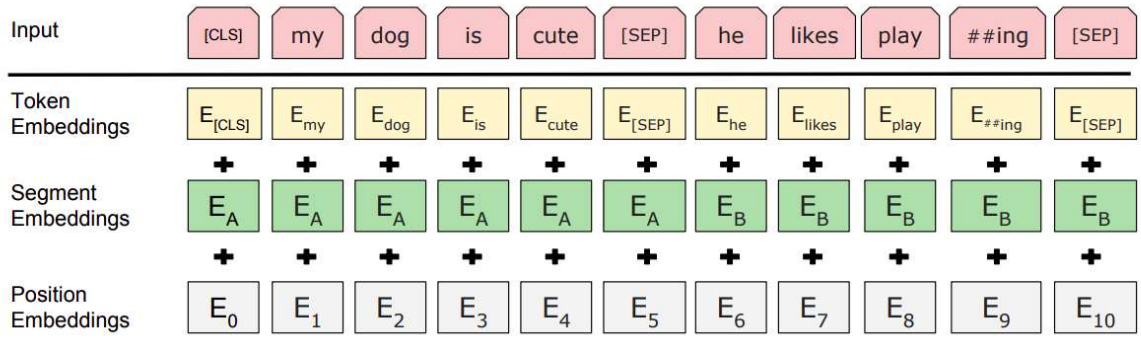

Figure 1: BERT input representation.

After pretraining on unlabeled data, BERT can be fine-tuned on various NLP tasks, such as text classification, information extraction, question-answering systems, and more. However, fine-tuning the model requires labeled data specific to the task.

BERT and its various variations, such as RoBERTa, ALBERT, and ELECTRA, have achieved high performance in many NLP tasks and have become a standard tool for research and applications. They provide a powerful tool for working with natural language and enable the extraction of rich semantic representations from textual data.

2.3.2. Key Elements of BERT Architecture

Word Embeddings: The input text is tokenized into individual words, called tokens. Each token is represented by a vector representation known as word embeddings. BERT utilizes word embeddings that can be pretrained or trained together with the model.

Segment Embeddings: When processing pairs of sentences, BERT uses special segment embeddings to separate and distinguish between the sentences. Each token is marked with a segment identifier, indicating its belonging to a specific sentence.

Positional Embeddings: To capture positional information in the text, BERT employs positional embeddings. They represent the relative positions of tokens within a sentence and allow the model to consider word order.

Multi-layer Transformers: The main component of the BERT architecture consists of several layers of transformers. Each layer comprises two sub-layers: a self-attention mechanism and a fully connected neural layer with activation function (e.g., ReLU). The transformer structure enables modeling of contextual dependencies between words in a sentence.

Pooling: To obtain a fixed-size representation of the entire sentence from the transformer outputs, BERT employs pooling. Typically, mean pooling is used, which computes the average value of all token outputs.

These elements, including word embeddings, segment embeddings, positional embeddings, multi-layer transformers, and pooling, work together to capture contextual information, preserve word order, and generate powerful representations for semantic text understanding in BERT.

3. Experiments and Results

A possible overall solution can be as follows:

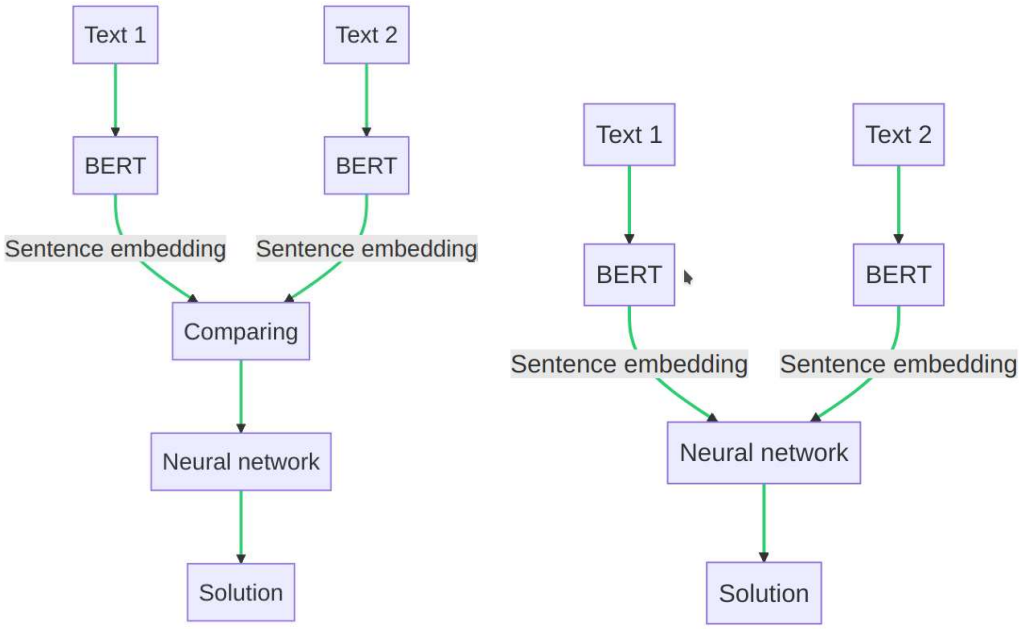

Figure 2. Schemes of possible solutions.

One approach is to obtain two vectors from two BERT heads and compare them using cosine or Euclidean distance. Then, train a neural network to interpret this distance more accurately.

Another approach is to train the neural network directly on the vectors themselves, rather than on the distances.

3.1. Initial Attempts to Build the Algorithm

For the initial experiments, a dataset with translated Russian sentences and the RuBERT model were used to obtain embeddings.

3.1.1. Obtaining Embeddings

The embeddings were taken from the very last layer of the neural network and then compared using cosine distances.

To assess the overall correctness of the model's performance with the data, a decision was made to study the distribution diagram of the target label against the cosine distance, as well as calculate the correlation obtained from it.

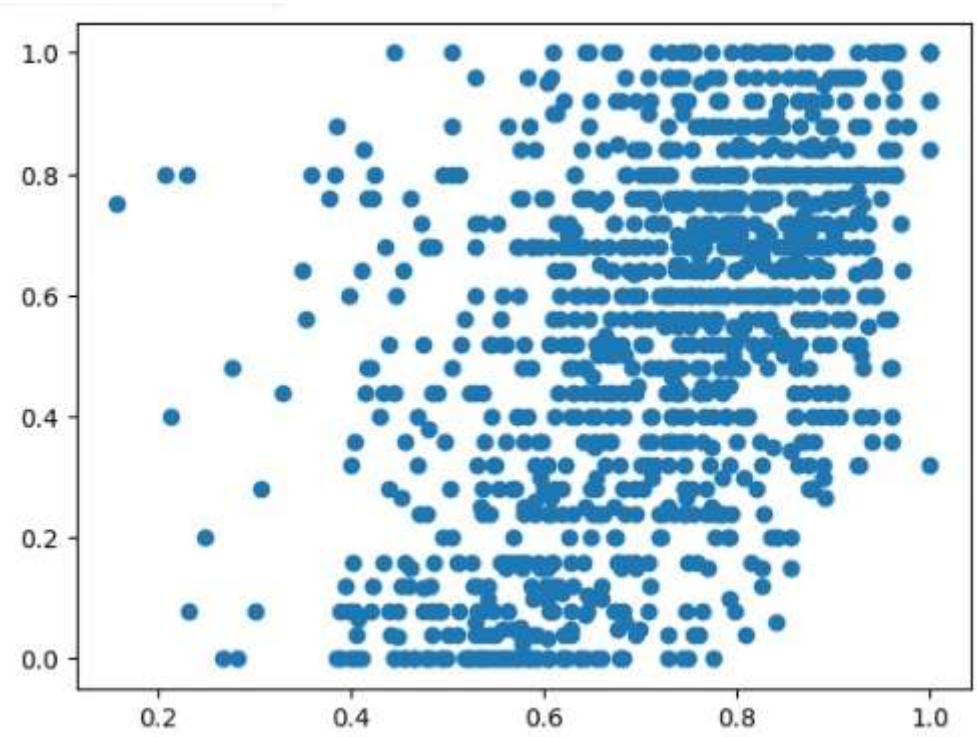

Figure 3. Diagram of the distribution of the label from the cosine distance.

The correlation between the target label and the cosine distance was found to be 0.484, which does not appear to be a good indicator. Nevertheless, a decision was made to train the neural network and evaluate the results obtained.

3.1.2. Building a Neural Network for Generating Results

A decision was made to develop a neural network that would determine semantic similarity based on Euclidean distances. This approach allows measuring the distance between representation vectors and using it as a metric for determining the degree of semantic similarity between texts.

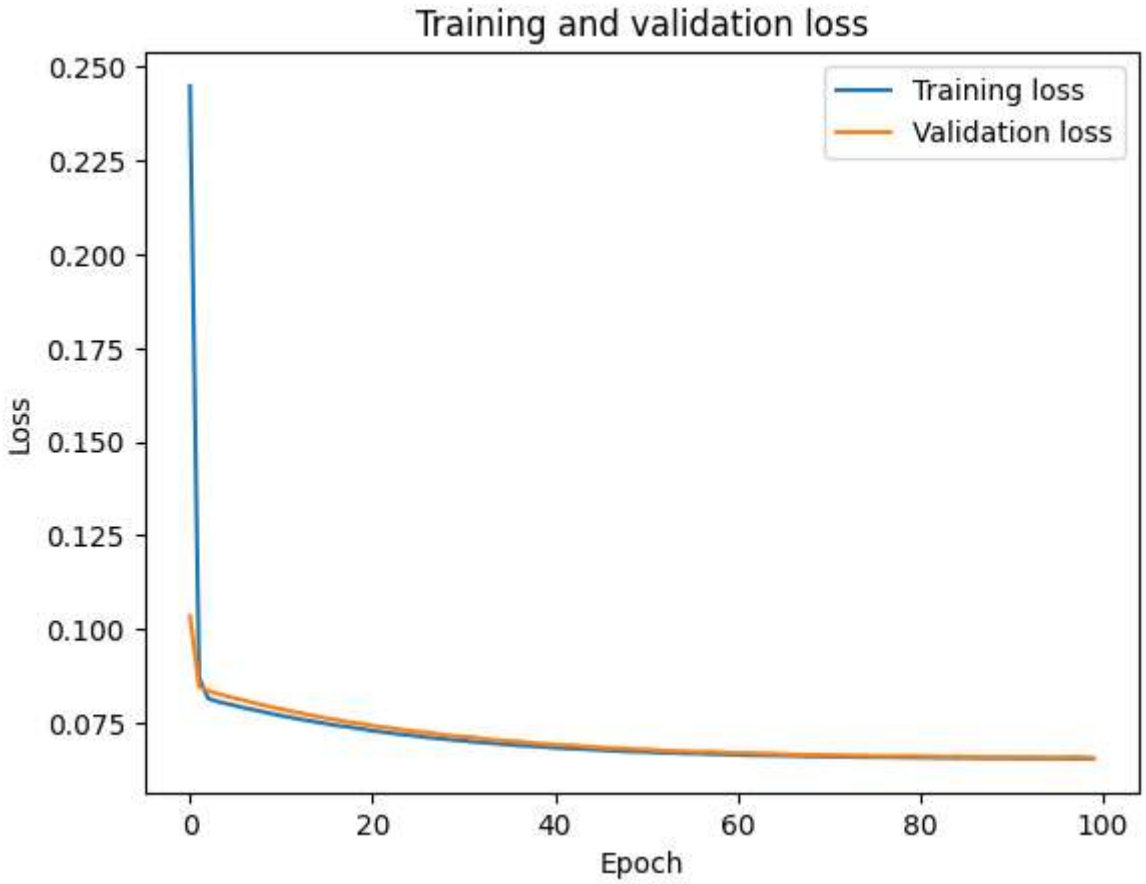

Figure 4. Graph of losses on test and training samples when training NS at cosine distances

The quality of the obtained algorithm can be measured by the mean squared error, which was found to be 0.0656.

3.2. Attempts to Improve the Model

Several approaches were explored to improve the model:

- Testing alternative BERT models.
- Switching to the original language (English).
- Converting cosine distances to Euclidean distances.

Subsequently, all experiments were conducted using a dataset in the original language.

3.2.1. Comparative Analysis of Pretrained BERT Models

To improve the model, one approach is to replace the model used for obtaining the embeddings. The following table provides an overview of different pretrained BERT models:

Table 1. Models for testings

Model Identifier	Model Name
m0	all-MiniLM-L6-v2
m1	distilbert-base-multilingual-cased
m2	sentence-transformers/distiluse-base-multilingual-cased-v1
m3	inkoziev/sbert_synonymy
m4	sentence-transformers/stsb-xlm-r-multilingual
m5	paraphrase-multilingual-MiniLM-L12-v2

After obtaining the embeddings, the model can be evaluated based on several criteria. The first criterion can be the correlation between the obtained sentence vectors and the dataset labels. This means that the model should create embeddings that have a high correlation with the true values or ratings in the dataset. A higher correlation indicates that the model accurately captures the semantic similarity between sentences.

The second criterion can be the dimensionality of the obtained embeddings. A good model should produce embeddings that preserve important semantic properties of texts while maintaining a reasonable size. Embeddings that are too large can lead to increased computational complexity and memory consumption.

By evaluating the models based on these criteria, the best model for the task of semantic text comparison can be selected. This will improve the quality of text representations and enhance the accuracy in determining their semantic similarity.

Table 2. Result of BERT-model comparisons.

	m0	m1	m2	m3	m4	m5
The value of the correlation with the label	-0.858	-0.570	-0.822	-0.635	-0.866	-0.86
Calculation time 2000 embeddings (s.)	18.62	19.81	21.26	9.42	38.44	34.41

During the comparative analysis of models for semantic text comparison, it was found that two models showed the best results. One of them is "paraphrase-multilingual-MiniLM-L12-v2," and the other is "sentence-transformers/stsb-xlm-r-multilingual."

Model m5, "paraphrase-multilingual-MiniLM-L12-v2," was chosen as the optimal option, despite its slightly lower correlation compared to "sentence-transformers/stsb-xlm-r-multilingual." It has a lower embedding dimensionality of 384. This improves computational speed, which is a significant advantage in processing large volumes of text data.

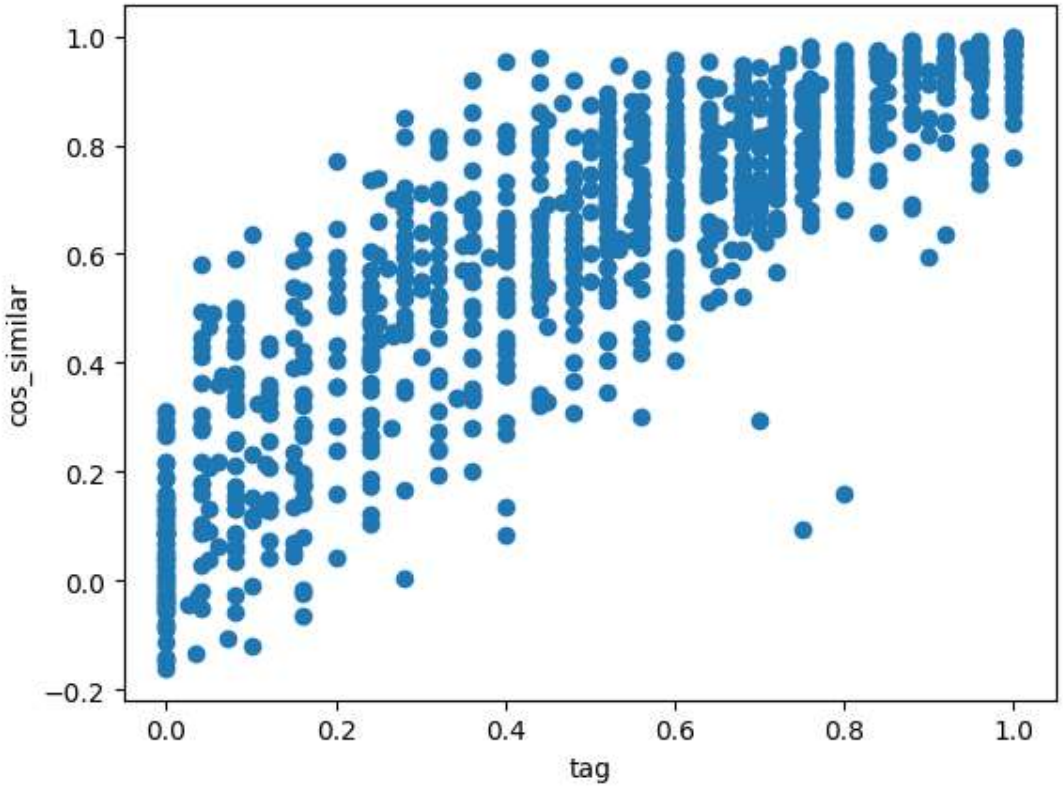

Figure 5. Distribution of the dependence of cosine distances on the target label

The experiments revealed that the correlation between the cosine distance, based on the embeddings of the "paraphrase-multilingual-MiniLM-L12-v2" model, and the target variable is 0.86. This indicates a strong relationship between the semantic similarity of texts determined by the model and the true values in the target variable. Such a high correlation confirms the effectiveness of the model in the task of semantic text comparison.

Thus, the "paraphrase-multilingual-MiniLM-L12-v2" model represents a promising solution for the task of semantic text comparison, thanks to its relatively lower embedding dimensionality and high correlation with the target variable.

3.2.2. Converting Cosine Distances to Euclidean Distances

During the experiment using cosine distance to measure semantic text similarity, it was observed that the obtained plot of cosine distance against the target variable might be unsatisfactory.

To improve the results, the idea of normalizing the vectors and using the Euclidean metric instead of cosine distance emerged.

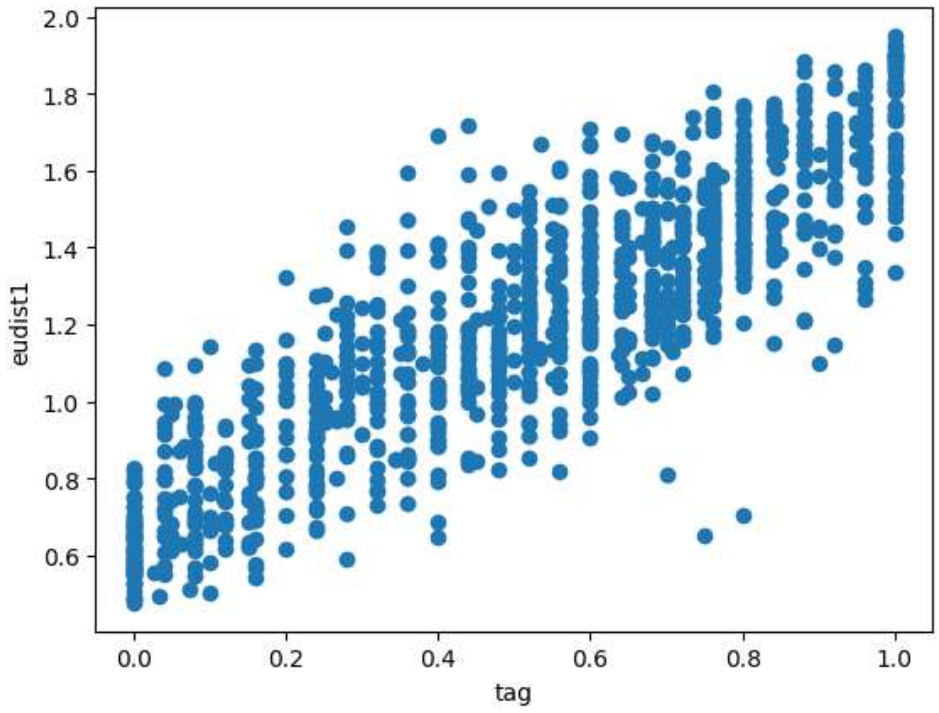

Figure 6. Distribution of the dependence of normalized Euclidean distances to a unit length from the target label

After applying vector normalization and transitioning to the Euclidean metric, it was found that the correlation value improved compared to the previous approach. Additionally, the plot of the Euclidean metric against the target variable was more satisfactory.

	tag	cos_dist	cos_similar	eudist1
tag	1.000000	-0.860156	0.860156	-0.868712
cos_dist	-0.860156	1.000000	-1.000000	0.974745
cos_similar	0.860156	-1.000000	1.000000	-0.974745
eudist1	-0.868712	0.974745	-0.974745	1.000000

Figure 7. Correlation matrix

Thus, using the Euclidean metric instead of cosine distance led to a higher correlation value and produced a more convincing and visually appealing plot depicting the relationship between the metric and the target variable.

This confirms the effectiveness of using the Euclidean metric for the task of semantic text comparison and may be a preferable option.

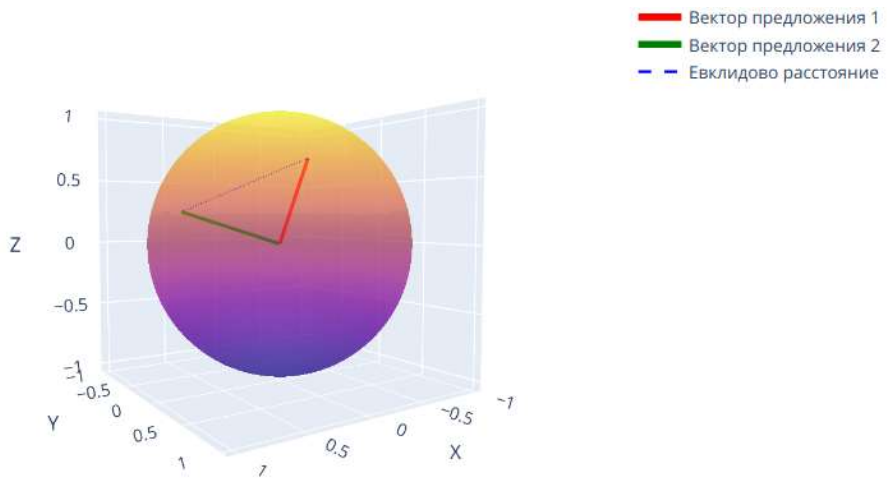

Figure 8. Visualization of the Euclidean distance for normalized vectors.

3.2.3. Building a Neural Network based on Embeddings

During the task of semantic text comparison, there was a need to choose between two approaches. The first approach involved training a neural network on reduced vectors obtained using dimensionality reduction methods. The second approach aimed to train a neural network on the distances between sentence vectors to determine the degree of their similarity.

Let's consider training a neural network based on the obtained vectors. To reduce the dimensionality of the vectors, the Principal Component Analysis (PCA) method was applied. Using this method, the dimensionality of the vectors was reduced to 30.

The neural network was trained on the reduced vectors using various topologies, hyperparameters, activation functions, and regularization methods. The network topologies varied from simple single-layer models to more complex deep architectures. ReLU was chosen as the activation function. Regularization was

applied in a combined manner, using either L1 or L2 regularization, to mitigate overfitting.

The results of training the neural network on the vectors showed a relatively high val_loss value of 0.05. This indicates that the model's ability to generalize and predict the semantic similarity of texts is not very strong.

3.2.4. Building a Neural Network based on Euclidean Distances

As an alternative approach, a neural network was trained using the Euclidean distances between pairs of sentence vectors.

Similarly, to the previous case, various hyperparameters were explored, including the network topology, activation functions, and regularization methods.

The results led to the selection of a simple neural network with two hidden layers containing 5 and 4 neurons, respectively. The ReLU activation function performed the best. No regularization methods were required. The Nesterov momentum optimizer was chosen.

After training the neural network for 1000 epochs using the Euclidean distances, a low val_loss value of approximately 0.019 was achieved. This indicates that the model is successfully learning to predict the similarity of sentences based on their Euclidean distances.

Conclusion

During this work, detailed research was conducted in the field of semantic text comparison using machine learning methods. An overview of existing methods and approaches to semantic text comparison was carried out. Special attention was given to the BERT architecture, which is one of the most powerful and flexible models for semantic text comparison. Various ways of using BERT for semantic text comparison were considered, including comparison based on cosine and Euclidean distances, as well as using a neural network trained on vectors.

As part of the work, a machine learning algorithm for semantic text comparison based on the BERT architecture was implemented. Experiments were conducted to optimize the algorithm and improve its performance. The results of the experiments showed that the proposed approach provides high accuracy in semantic text comparison and can be successfully applied in a range of tasks related to natural language processing.

Overall, the results of this work confirm the relevance and importance of applying machine learning methods to semantic text comparison. They also highlight the potential of the BERT architecture as one of the most powerful tools in this field. In the future, these results can be used to develop more advanced natural language processing systems capable of effectively analyzing and comparing texts at the semantic level.