

Школа Инженерная школа информационных систем и технологий
 Направление подготовки Информационные системы и технологии
 Отделение школы Информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Разработка программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей

УДК 004.415:82.05:808.1:004.032.26

Студент

Группа	ФИО	Подпись	Дата
8И5Б	Михайлова Елена Владиславна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения ИТ	Иванова Юлия Александровна	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОСГН ШБИП	Шулинина Юлия Игоревна			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОБД	Немцова Ольга Александровна			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения ИТ	Цапко Ирина Валериевна	к.т.н.		

Планируемые результаты обучения по ООП

Код результатов	Результат обучения (выпускник должен быть готов)
	<i>Профессиональные и общепрофессиональные компетенции</i>
Р1	Применять базовые и специальные естественнонаучные и математические знания для комплексной инженерной деятельности по созданию, внедрению и эксплуатации геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе.
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием геоинформационных систем и технологий, информационных систем в бизнесе, с использованием базовых и специальных знаний, современных аналитических методов и моделей.
Р4	Выполнять комплексные инженерные проекты по созданию информационных систем и технологий, а также средств их реализации (информационных, методических, математических, алгоритмических, технических и программных).
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе.
Р6	Внедрять, эксплуатировать и обслуживать современные геоинформационные системы и технологии, информационные системы и технологии в бизнесе, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.

Министерство образования и науки Российской Федерации

федеральное государственное автономное образовательное учреждение
высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа Инженерная школа информационных технологий и робототехники

Направление подготовки 09.03.02 Информационные системы и технологии

Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:

Руководитель ООП

(Подпись) (Дата) Цапко И.В.
(Ф.И.О.)

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

В форме:

Бакалаврская работа

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

Студенту:

Группа	ФИО
8И5Б	Михайловой Елене Владиславне

Тема работы:

Разработка программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей	
Утверждена приказом директора (дата, номер)	18.02.2019 г. № 1240/с

Срок сдачи студентом выполненной работы:	03.06.2019
--	------------

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе <i>(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.).</i>	Работа направлена на создание программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей.
--	--

Перечень подлежащих исследованию, проектированию и разработке вопросов <i>(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).</i>	Анализ предметной области; проектирование программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей; изучение алгоритмов и моделей искусственных нейронных сетей; реализация программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей; выполнение численных экспериментов; финансовый менеджмент; социальная ответственность.
Перечень графического материала	
Консультанты по разделам выпускной квалификационной работы	
Раздел	Консультант
Финансовый менеджмент	Шулинина Юлия Игоревна
Социальная ответственность	Немцова Ольга Александровна
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Все разделы должны быть написаны на русском языке	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	21.01.2019
--	-------------------

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения ИТ	Иванова Юлия Александровна	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5Б	Михайлова Елена Владиславна		

**ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И
РЕСУРСОСБЕРЕЖЕНИЕ»**

Студенту:

Группа	ФИО
8И5Б	Михайловой Елене Владиславне

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	бакалавриат	Направление/специальность	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. <i>Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих</i>	Оклад инженера – 21760 руб. Оклад руководителя – 33664 руб.
2. <i>Нормы и нормативы расходования ресурсов</i>	Премияльный коэффициент 30%; Коэффициент доплат и надбавок 20%; Районный коэффициент 30%; Коэффициент дополнительной заработной платы 15%; Накладные расходы 16%.
3. <i>Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования</i>	Коэффициент отчислений на уплату во внебюджетные фонды 30%

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. <i>Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения</i>	-Анализ конкурентных технических решений
2. <i>Планирование и формирование бюджета научных исследований</i>	Формирование плана и графика разработки: - определение структуры работ; - определение трудоемкости работ; - разработка графика Гантта. Формирование бюджета затрат на научное исследование: - материальные затраты; - затраты на специальное оборудование; - заработная плата (основная и дополнительная); - отчисления на социальные цели; - накладные расходы.
3. <i>Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования</i>	- Определение потенциального эффекта исследования

Перечень графического материала (с точным указанием обязательных чертежей):

1. <i>Оценочная карта конкурентных технических решений</i>
2. <i>Матрица SWOT</i>
3. <i>График Гантта</i>
4. <i>Расчет бюджета затрат</i>

Дата выдачи задания для раздела по линейному графику

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОСГН ШБИП	Шулинина Ю.И.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5Б	Михайлова Е.В.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8И5Б	Михайловой Елене Владиславне

Школа		Отделение (НОЦ)	
Уровень образования	бакалавриат	Направление/специальность	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Социальная ответственность»:

1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Работа направлена на создание программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– Специальные правовые нормы трудового законодательства при работе с компьютером и орг. техникой; – требования к организации рабочих мест пользователей. – гигиенические требования к персональным электронно-вычислительным машинам и организации работы СанПиН 2.2.2/2.4.1340-03; – для соблюдения всех санитарно-технических и гигиенических правил и требований рассмотрен Трудовой кодекс; – гигиенические требования к видео-дисплейным терминалам, персональным электронно-вычислительным машинам и организации работы СанПиН 2.2.542-96
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	В качестве вредных и опасных факторов выделены: – повышенный уровень шума на рабочем месте; – повышенный уровень электромагнитных излучений; – перенапряжение зрительного анализатора; – отклонение показателей микроклимата; – недостаточная освещенность рабочей зоны;
3. Экологическая безопасность:	– Влияния на гидросферу и атмосферу разработка не оказывает; – анализ воздействия на литосферу (в работе рассмотрена утилизация неисправного ПК).
4. Безопасность в чрезвычайных ситуациях:	– Основным ЧС при работе в офисном помещении является пожар.

Дата выдачи задания для раздела по линейному графику	
---	--

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент	Немцова Ольга Александровна			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5Б	Михайлова Елена Владиславна		

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

Школа Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.03.02 Информационные системы и технологии
Уровень образования бакалавриат
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2018/2019 учебного года

Форма представления работы:

Бакалаврская работа

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

**КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы**

Срок сдачи студентом выполненной работы:	03.06.2019
--	------------

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
21.02.2019	Анализ предметной области	20
21.03.2019	Проектирование программного обеспечения	20
27.05.2019	Реализация программного обеспечения	35
03.06.2019	Финансовый менеджмент	15
03.06.2019	Социальная ответственность	10

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения ИТ	Иванова Юлия Александровна	к.т.н.		

СОГЛАСОВАНО:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения ИТ	Цапко Ирина Валериевна	к.т.н.		

РЕФЕРАТ

Выпускная квалификационная работа 80 с., 30 рис., 19 табл., 23 источника.

Ключевые слова: ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ, LSTM, СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ, WORD2VEC, ОПРЕДЕЛЕНИЕ АВТОРСТВА, ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ.

Объектом исследования являются нейросетевые методы классификации.

Цель работы: разработать программное обеспечение для определения автора по фрагменту произведения на основе искусственных нейронных сетей.

В процессе исследования проводились: анализ предметной области, изучение моделей искусственных нейронных сетей, конфигурация, обучение и тестирование сверточных и рекуррентных нейронных сетей.

В результате исследования: разработано программное обеспечение для определения автора по фрагменту произведения на основе искусственных нейронных сетей.

Степень внедрения: выполнено в рамках проекта «Создание интеллектуальной системы детектирования, распознавания и понимания искаженных печатных текстов на изображениях и видео», поддержанный грантом РФФИ № 18-08-0097а.

Область применения: разработанное программное обеспечение может применяться при проверке письменных работ на наличие плагиата, а также для определения авторства текста.

Оглавление

Введение.....	11
1 Анализ предметной области определения автора текста.....	13
1.1 Описание существующих решений определения автора текста.....	13
1.1.1 Система «Лингвоанализатор»	13
1.1.2 Система «Атрибутор»	14
1.1.3 Система «Авторовед».....	14
1.2 Сверточные нейронные сети и их архитектура	15
1.2.1 Полносвязный слой	16
1.2.2 Сверточный слой	17
1.2.3 Субдискретизирующий слой	17
1.3 Рекуррентные нейронные сети	18
1.3.1 Архитектура рекуррентных нейронных сетей.....	18
1.3.2 LSTM рекуррентные нейронные сети и их архитектура.....	20
1.4 Описание технологии Word2Vec.....	24
1.5 Цели и задачи работы	25
2 Проектирование и реализация программного обеспечения	26
2.1 Обоснование выбора программного обеспечения.....	26
2.2 Проектирование программного обеспечения для определения авторства.....	26
2.2.1 Реализация сверточной нейронной сети	28
2.2.2 Проектирование и реализация рекуррентной нейронной сети для определения авторства	29
2.2.1.1 Архитектура рекуррентной нейронной сети	29
2.2.1.2 Реализация рекуррентной нейронной сети.....	30
3 Обучение и тестирование	32
3.1 Задача классификация текстов	32
3.1.1 Настройка сверточной нейронной сети для классификации текстов	32
3.1.1.1 Описание исходной выборки для задачи классификации текстов...	32
3.1.1.2 Описание формирования модели Word2Vec и обработка выборки.	33

3.1.1.3 Обучение и тестирование сверточной нейронной сети для классификации текста.....	34
3.1.1.4 Результаты обучения и тестирования сверточной нейронной сети для классификации текстов.....	37
3.2 Задача определения авторства.....	37
3.2.1 Настройка сверточной нейронной сети для определения автора текста	38
3.2.1.1 Описание исходной выборки.....	38
3.2.1.2 Описание формирования модели Word2Vec и обработка выборки.	38
3.2.1.3 Обучение и тестирование сверточной нейронной сети для определения авторства.....	40
3.2.1.4 Результаты обучения и тестирования сверточной нейронной сети для определения авторства.....	43
3.2.2 Настройка рекуррентной нейронной сети для определения автора текста	44
3.2.1.1 Описание формирования модели Word2Vec и обработка выборки.	44
3.2.1.2 Обучение и тестирование рекуррентной LSTM нейронной сети.....	45
3.2.1.3 Результаты обучения и тестирования рекуррентной LSTM нейронной сети	45
3.2.1.4 Результаты обучения и тестирования рекуррентной LSTM нейронной сети с предварительной классификацией текстов	46
3.2.3 Сравнение результатов сверточной и рекуррентной нейронной сети для определения авторства	47
4 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ.....	48
5 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ.....	64
Заключение	76
Список использованных источников и литературы	78

Введение

Автороведческая экспертиза — исследование печатного текста для установления его автора (задача атрибуции произведения) или выяснения каких-либо сведений об условиях создания текстового документа [1].

Задача атрибуции встречается в различных областях, например, в литературоведении, музееведении, при ведении архивов, и представляет интерес для филологов, литературоведов, юристов, криминалистов, историков. Для решения задачи определения авторства зачастую приходится обращаться к экспертам, которые при помощи характерных стилистических приемов или языковых особенностей могут идентифицировать автора неизвестного текста. Однако, экспертный анализ авторского стиля является трудоемким процессом. Поэтому появляется потребность в создании систем, способных автоматизировать этот процесс [2].

Все вышесказанное делает актуальной задачу определения автора по фрагменту текста.

Целью работы является разработка программного обеспечения для определения автора по фрагменту произведения на основе искусственных нейронных сетей.

Объектом исследования являются нейросетевые методы классификации. Предметом исследования является автоматизация процесса определения авторства по фрагменту произведения.

В первой главе приведен анализ предметной области, рассмотрены существующие решения, описаны технологии и алгоритмы решения задачи определения авторства.

Во второй главе приведено обоснование выбора программного обеспечения; описан процесс проектирования программного обеспечения, проектирования и реализации его компонентов.

В третьей главе приведено описание обучения и тестирования нейронных сетей для задачи классификации текста и определения авторства.

Описана исходная выборка, формирование модели Word2Vec и обработка выборки с её помощью. Описан подбор параметров обучения для каждой из нейронных сетей, приведены результаты работы сетей и их сравнение. Выбран наилучший способ решения задачи определения авторства.

В четвертой главе изложены вопросы, связанные с финансовым менеджментом и ресурсоэффективностью проекта.

В пятой главе рассмотрены экологическая безопасность и техника безопасности работников на рабочем месте.

Работа выполнена в рамках проекта «Создание интеллектуальной системы детектирования, распознавания и понимания искаженных печатных текстов на изображениях и видео», поддержанный грантом РФФИ № 18-08-0097а.

1 Анализ предметной области определения автора текста

В качестве инструментов для определения авторства текста могут быть использованы как искусственные нейронные сети, так и другие математические модели. Далее будут рассмотрены существующие решения задачи определения автора по фрагменту текста.

1.1 Описание существующих решений определения автора текста

1.1.1 Система «Лингвоанализатор»

Система «Лингвоанализатор» была разработана Д.В. Хмелевым. Метод решения задачи, предложенный Хмелевым, основывается на использовании такого алгоритма как цепь А.А. Маркова. Алгоритм предполагает вычисление матрицы переходов. В данном случае, для задачи определения авторства, матрица переходных частот употреблений биграмм вычисляется только по тем произведениям автора, которые достоверно им созданы. Вычисленная матрица необходима для оценки матрицы вероятностей перехода из буквы в букву. Для каждого автора в базе системы строится матрица. Также для каждого автора оценивается вероятность того, что именно он написал анонимный фрагмент текста. Система полагает, что автором анонимного литературного произведения является, у которого рассчитанная оценка вероятности выше [3].

Кроме того, в данной системе используется метод, основанный на применении относительной энтропии, при котором нет необходимости обрабатывать исходный текст. Этот метод основан на сжатии текста. При необходимости определить авторство неизвестного фрагмента литературного произведения, он добавляется к тексту, автор которого известен. Далее происходит сжатие полученного фрагмента. Система полагает, что правильный исходный класс документа тот, на котором он жмется лучше всего [4].

Апробация системы проводилась на произведениях 82 писателей. Точность метода составила 84%.

1.1.2 Система «Атрибутор»

Программный продукт «Атрибутор» представляет из себя систему для сравнения текстов при помощи машинного обучения, а также их классификации по характеристикам авторского стиля.

Система основана на методе, использующем триграммы, то есть трехбуквенные сочетания, в качестве стилевых признаков автора. В таком случае, в анализе учитываются одно- и двухбуквенные сочетания, например, предлоги или другие служебные слова, частота употребления которых считается характерным стилистическим показателем.

В базе системы содержится различные литературные произведения 103 авторов, из которых исключены имена собственные. Литературные произведения выбирались таким образом, что тексты разных писателей имели максимальное число различий, а тексты одного и того же автора различались минимально. Такой подбор произведений увеличивает разрыв между различными классами текстов и повышает вероятность успешного определения авторства. Обучающую выборку системы формируют романы отечественных писателей 19 – 20 века. Длина каждого текста должна была составлять не менее шести печатных страниц. Это связано с тем, что небольшие фрагменты текстов могут оказаться не показательными для выявления стилистических особенностей автора и вычисления авторской инварианты.

1.1.3 Система «Авторовед»

Программное решение основано на нейронных сетях. Оно составляет психологический портрет автора, основываясь на анализе фонетики, лексики и семантики текста.

Система использует метод опорных векторов для формирования обучающей выборки. Идея метода заключается в том, чтобы перевести исходные фрагменты литературных произведений, загруженных в базу системы в вектора, а далее в пространство более высокой размерности с

дальнейшим поиском гиперплоскости, максимально разделяющей тексты на классы по авторам.

Метод работает в том случае, когда полагается, что максимальное расстояние между классами текстов обеспечит минимальную среднюю ошибку классификатора [5]. Далее, после разделения исходных текстов на классы, для классификации фрагмента неизвестного литературного произведения достаточно проверить принадлежность текста одному из пространств.

Также в системе «Авторовед» используются искусственные нейронные сети. В ее основе лежат многослойный перцептрон и сети каскадной корреляции. База системы состоит из 215 текстов 50 отечественных писателей. Размер каждого текста составлял более 100 000 символов. Обучающая выборка для каждого автора – 3 текста. Тестовая – 1 текст. Точность системы составляет 95 %.

Исходя из анализа существующих систем, можно сделать вывод, что применение искусственных нейронных сетей предоставляет наиболее точное решение задачи определения авторства по фрагменту текста. В связи с этим, в данной работе для решения задачи определения авторства текста было решено руководствоваться нейросетевым подходом.

1.2 Сверточные нейронные сети и их архитектура

Обычно архитектура сверточной нейронной сети представляет собой чередование сверточных слоев (также известных как convolution layers), осуществляющих операцию свертки, субдискретизирующих слоев (subsampling layers), уменьшающих размерность исходной выборки и полносвязных слоев (fully-connected layer) на выходе. Количество нейронов в полносвязном слое выходного слоя определяет количество выходных классов сети. Все три вида слоев могут чередоваться в произвольном порядке [6].

В сверточном слое нейроны, у которых в ходе обучения и тестирования используются одинаковые значения весовых коэффициентов,

объединяются в карты признаков (feature maps), а каждый нейрон карты признаков связан с частью нейронов предыдущего слоя. Таким образом, при вычислении сети каждый нейрон сверточного слоя выполняет свертку некоторой области предыдущего слоя, которая определяется множеством нейронов, связанных с данным нейроном. На рисунке 1 представлен пример архитектуры сверточной нейронной сети.

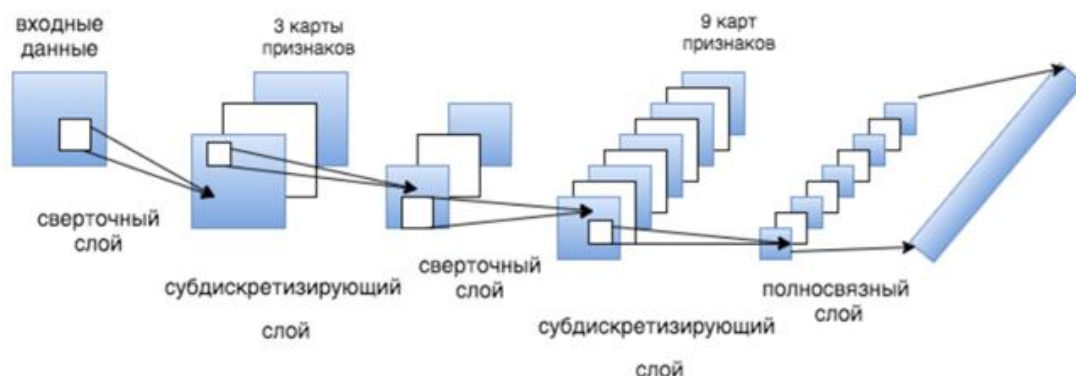


Рисунок 1 – Архитектура сверточной нейронной сети

Из рисунка 1 видно, что на вход нейронной сети подается набор входных данных. Первый слой сети сверточный, в процессе его обучения происходит разделение нейронов на 3 карты признаков. Далее на субдискретизирующем слое происходит уменьшение размерности. В течении обучения количество карт признаков может меняться. Чередование сверточных и субдискретизирующих слоев обеспечивает выделение признаков из исходных данных.

1.2.1 Полносвязный слой

Полносвязным слоем называют слой, в котором каждый нейрон текущего слоя связан со всеми нейронами предыдущего, причем каждой связи присвоен свой весовой коэффициент. На Рисунок 2 приведен пример полносвязного слоя, где x_k – входные сигналы, w_{ij} – весовые коэффициенты, f – функция активации.

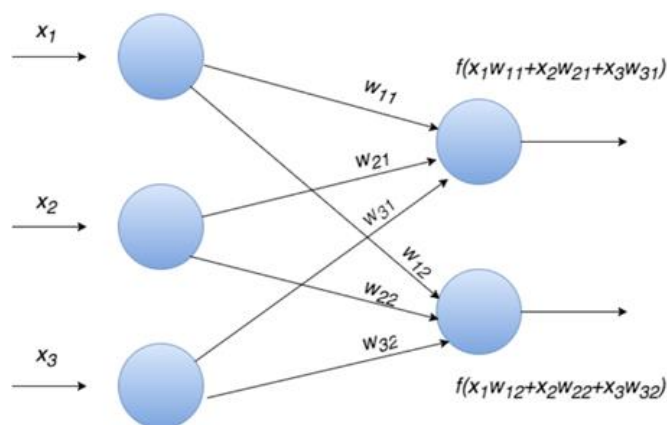


Рисунок 2 – Полносвязный слой нейронной сети

1.2.2 Сверточный слой

В сверточном слое, в отличие от полносвязного, нейрон соединен лишь с ограниченным количеством нейронов предыдущего уровня. Таким образом, сверточный слой применяет к исходным данным операцию свертки – используется лишь матрица весовых коэффициентов небольшого размера (ядро свертки), которую «передвигают» по всему обрабатываемому слою. За счет так называемых краевых эффектов сверточный слой немного уменьшает размерность исходных данных. На Рисунок 3 представлен пример сверточного слоя с ядром свертки размера 3×3 .

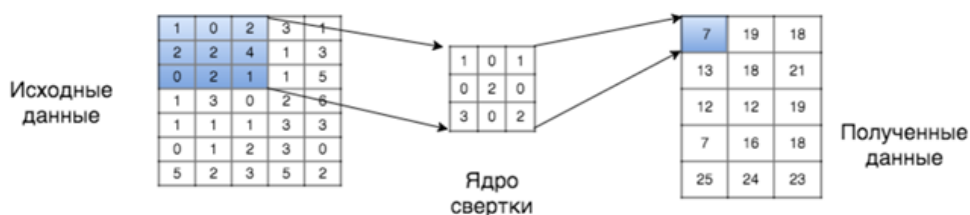


Рисунок 3 – Сверточный слой

1.2.3 Субдискретизирующий слой

Слои этого типа отвечают за уменьшение размерности исходных данных (обычно в несколько раз, в зависимости от способа и параметров субдискретизации). Субдискретизацию можно делать разными способами, например, при помощи выбора максимального или вычисления среднего значения по ячейкам. Чаще всего используется метод выбора максимального

элемента, так называемый max-pooling. При этом способе вся карта признаков разделяется на ячейки, из которых выбираются максимальные по значению, как показано на рисунке 4. Таким образом происходит уменьшение размерности исходных данных.



Рисунок 4 – Субдискретизирующий слой

1.3 Рекуррентные нейронные сети

1.3.1 Архитектура рекуррентных нейронных сетей

Рекуррентные нейронные сети – это вид нейронных сетей, в котором связи между элементами сети образуют направленную последовательность [7].

В настоящее время рекуррентные нейронные сети успешно используются повсеместно, например, для задач распознавания речи, языкового моделирования, перевода, создания субтитров к изображениям.

Этот вид сетей отличаются от других тем, что в них каждый нейрон взаимодействует сам с собой, то есть сигнал предыдущего выхода передается еще раз на вход того же нейрона, формируя краткосрочную память. На вход сети передаётся сигнал, являющийся некоторой последовательностью. Каждый элемент последовательности поочерёдно передаётся одним и тем же нейронам, которые своё же предсказание возвращают себе вместе со следующим её элементом входной последовательности, до тех пор, пока эта последовательность не закончится. Использование рекуррентных нейронных сетей эффективно в областях, где исходная информация содержит последовательность (тексты, аудиосигналы). Элементы рекуррентной сети изображают как обычные нейроны с дополнительной циклической стрелкой, которая демонстрирует то, что кроме входного сигнала нейрон использует

также своё дополнительное скрытое состояние, как представлено на рисунке 5. Если «развернуть» это изображение, то получим функционирование нейрона во времени [8].

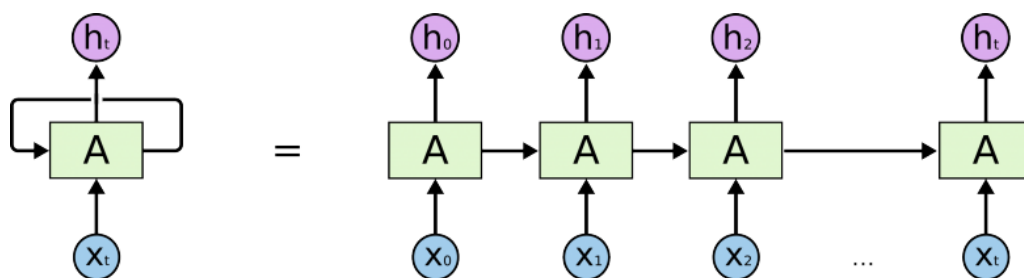


Рисунок 5 – Архитектурное представление рекуррентной нейронной сети

Такая архитектура нейронной сети позволяет решать различные задачи, например, предсказание последнего слова в предложении.

Моделирование памяти в нейронной сети подобным образом вводит новое измерение в описание процесса её работы — время. Пусть нейронная сеть получает на вход последовательность данных, например, текст пословно или слово по буквам. Тогда нейрон получает на вход каждый последующий элемент исходной последовательности в новый условный момент времени. К этому моменту в нейроне уже есть накопленный с начала поступления информации опыт, который состоит из выхода предыдущего нейрона.

Существуют различные типы рекуррентных нейронов, разделяющиеся между собой способами обработки ячейки памяти внутри них. Традиционно нейрон получает на вход некий исходный сигнал, применяет функцию активации (гиперболический тангенс, сигмоида и другие), передает полученный сигнал на вход следующего нейрона. Последующие нейроны получают на вход вектор памяти (выходной сигнал предыдущего нейрона) и сигнал из исходной последовательности. Получается обычная сетка с одним скрытым слоем. На рисунке 6 представлена схема описанной выше нейронной сети.

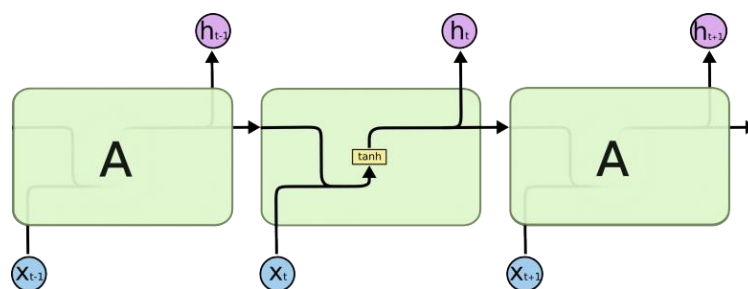


Рисунок 6 – Архитектурное представление рекуррентной нейронной сети с одним внутренним слоем

Однако память, реализованная подобным образом, получается весьма короткой. Поскольку каждый раз информация в памяти смешивается с информацией в новом сигнале, спустя некоторое небольшое количество итераций, как правило от 5 до 7, информация уже полностью перезаписывается. При решении задачи предсказания слова в предложении в пределах одного предложения такая рекуррентная нейронная сеть будет работать неплохо. Однако, в случае исходных текстов большого объема память сети будет перезаписываться многократно и закономерности в начале текста не будут оказывать влияния на предсказание сети и уже не будут вносить какой-либо вклад в решения сети ближе к концу текста, также как ошибка на первых элементах последовательностей в процессе обучения перестаёт вносить вклад в общую ошибку сети.

1.3.2 LSTM рекуррентные нейронные сети и их архитектура

Чтобы побороть недостаток с перезаписью памяти, была придумана LSTM-RNN сеть (Long Short-Term Memory Recurent Neural Network), в которой были добавлены дополнительные внутренние преобразования, которые позволяют дольше хранить полезную часть сигнала [9]. На рисунке 7 представлена схема LSTM-сети.

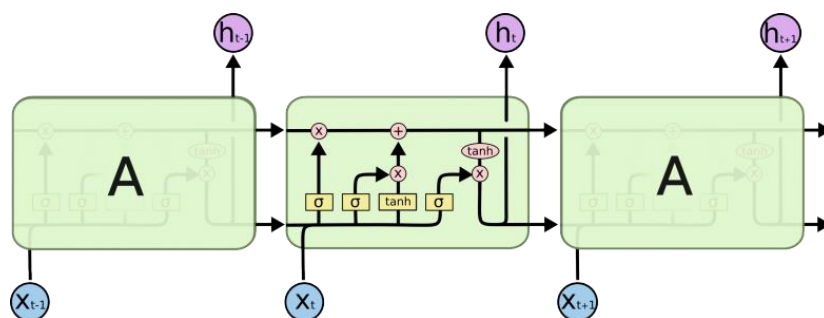


Рисунок 7 – Архитектурное представление LSTM сети с четырьмя взаимодействующими слоями

На рисунке 8 представлены используемые в вышеприведенной схеме обозначения.

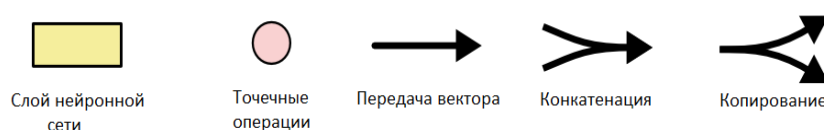


Рисунок 8 – Используемые обозначения

На диаграмме, приведенной на рисунке 7, каждая линия несет вектор от выхода одного узла до входа в другой. Круги представляют собой точечные операции, например, сложение векторов, а прямоугольники – обученные слои нейронной сети. Сливающиеся линии обозначают конкатенацию, а разветвляющиеся – копирование.

Основным понятием в работе LSTM сети является состояние ячейки (cell state, обозначается C_{t-1}), которая представляет собой вектор, представленный в виде горизонтальной линии на рисунке 9. Состояние ячейки можно сравнить с лентой конвейера, она проходит через всю сеть, участвуя лишь в нескольких преобразованиях.

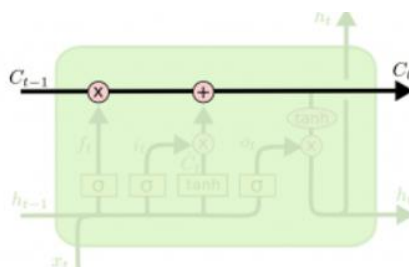


Рисунок 9 – Состояние ячейки LSTM сети

Однако, LSTM сеть может удалять информацию из состояния ячейки с помощью специальных настраиваемых структур, называемых фильтрами,

вентилими или вратами (gates). Фильтр – это структура, позволяющая пропускать информацию частично или полностью в зависимости от условий. Это структура состоит из слоя нейронной сети, имеющего сигмоидальную функцию активации и операции поточечного умножения. Пример фильтра представлен на рисунке 10.

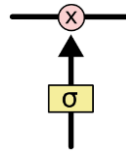


Рисунок 10 – Фильтр LSTM сети

Слой с сигмоидальной функцией активации в свою очередь возвращает числа от 0 до 1, которые обозначают какую часть исходной информации следует пропустить по сети дальше.

Рассмотрим LSTM сеть, изображенную на рисунке 7 послойно.

Первый слой, называемый слоем фильтра забывания (forget gate later) вычисляет с помощью сигмоиды, какую часть информации необходимо отбросить из состояния ячейки. Получая на вход h_{t-1} и x_t слой фильтра забывания f_t выдает на выходе число W_f от 0 до 1 для каждого значения состояния ячейки, как показано на рисунке 11.

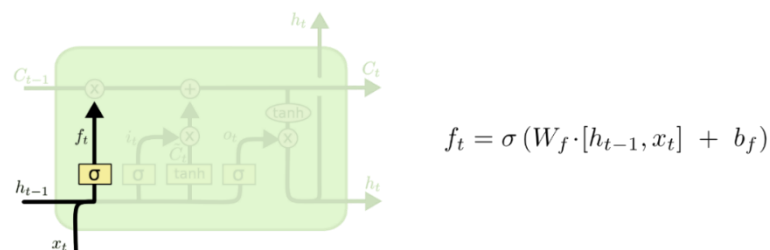


Рисунок 11 – Первый слой LSTM сети

Второй слой, называемый слоем входного фильтра (input layer gate), обозначаемый i_t , также с помощью сигмоиды вычисляет какую долю поступившей информации, пришедшей от выхода предыдущего нейрона h_{t-1} и новой исходной информации x_t , необходимо оставить и передать дальше по сети. Далее слой с функцией гиперболического тангенса составляет вектор новых значений-кандидатов $\tilde{C}_t$, которые можно добавить в состояние ячейки, как показано на рисунке 12.

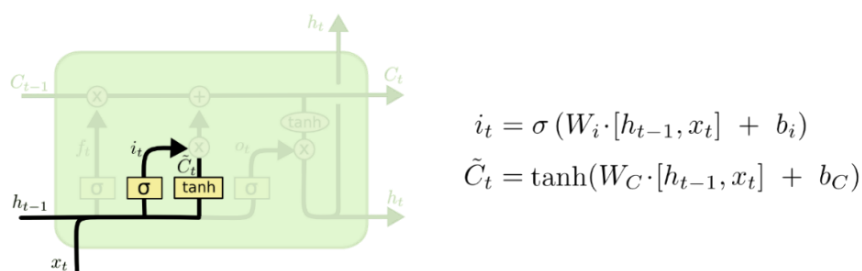


Рисунок 12 – Второй слой LSTM сети

На третьем слое сети вычисляется линейная комбинация старого сигнала и обновленного сигнала с только вычисленными весовыми коэффициентами для каждой из компонент. Так получается новое состояние ячейки C_t , которое в таком же виде передаётся далее, как показано на рисунке 13.

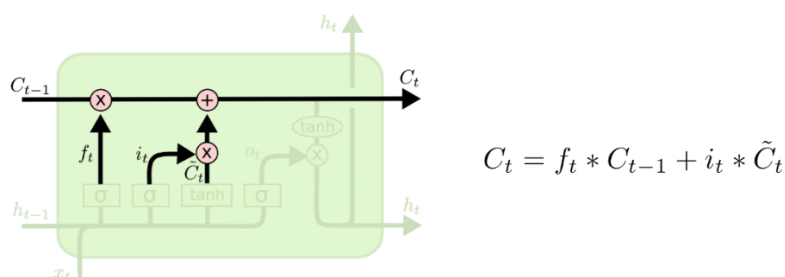


Рисунок 13 – Третий слой LSTM сети

В заключение необходимо вычислить выходной сигнал h_t . Так как часть входного сигнала уже в памяти, то не нужно считать активацию по всему сигналу. Сначала сигнал проходит через сигмоиду, которая определяет, какая его часть важна для дальнейших решений и получаем вектор o_t , затем гиперболический тангенс нормирует состояние ячейки C_t в диапазоне от -1 до 1, и в конце эти два вектора перемножаются, как показано на рисунке 14.

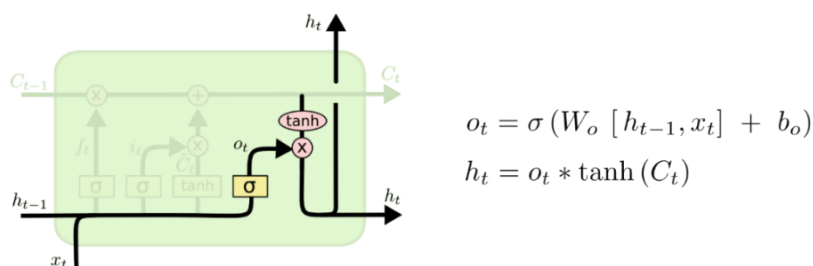


Рисунок 14 – Четвертый слой LSTM сети

Полученные таким образом h_t и C_t передаются далее по цепочке.

Существует множество вариантов используемых каждым слоем функции активации, возможны некоторые небольшие изменения самой схемы и каких-либо её параметров, однако, суть функционирования не меняется — сначала фильтруют часть памяти, затем запоминают часть нового сигнала, а уже потом на основе этих данных вычисляется результат.

Например, один из популярных вариантов LSTM сети представлен на рисунке 15. Его предложили в 2000 году Феликс Герс и Юрген Шмидхубер в своей научной работе [10].

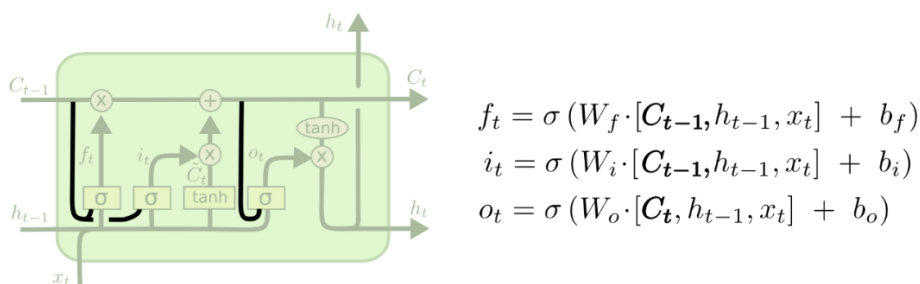


Рисунок 15 – LSTM сеть Герса и Шмидхубера

1.4 Описание технологии Word2Vec

Нейронные сети способны работать только с числовой информацией. Для того, чтобы представить текст в цифровом виде существует решение, называемое Word2Vec.

Word2vec – инструмент от компании Google для анализа семантики естественных языков и расчета векторных представлений слов [11].

Алгоритм работы технологии Word2Vec выглядит следующим образом: на вход технологии подается большой корпус текста; каждому слову входных данных сопоставляется вектор; на выходе Word2Vec выдает набор векторов слов. Для этого сначала он создает словарь, «обучаясь» на входных текстовых данных, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами, то есть близкие по значению, в векторном представлении будут иметь близкие вектора. В основе Word2Vec лежат нейронные сети. Возможно применение

двух архитектур: CBOW (Continuous Bag of Words) и Skip-gram. CBOW — «непрерывный мешок слов» модельная архитектура, которая предсказывает текущее слово, исходя из окружающего его контекста. Архитектура типа Skip-gram, приведенная на рисунке 16, действует иначе: она использует текущее слово, чтобы предугадывать окружающие его слова.

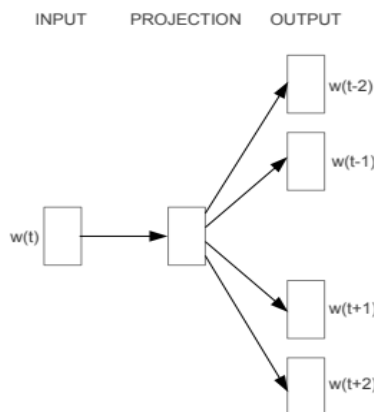


Рисунок 16 – Архитектура модели Skip-gram

Ни в одном из алгоритмов порядок слов не оказывает влияния на результат [12].

1.5 Цели и задачи работы

Исходя из анализа предметной области можно выделить следующую цель и задачи работы.

Цель работы: разработать программное обеспечение для определения автора по фрагменту произведения на основе искусственных нейронных сетей.

Задачи работы:

- изучить литературу и проанализировать полученные сведения;
- выбрать нейросетевые модели для реализации;
- спроектировать архитектуру программного обеспечения;
- спроектировать структуры выбранных нейросетевых моделей;
- обучить выбранные нейросетевых моделей;
- провести численные эксперименты на выборке при разных параметрах.

2 Проектирование и реализация программного обеспечения

2.1 Обоснование выбора программного обеспечения

В качестве библиотеки для машинного обучения была выбрана библиотека TensorFlow [13]. При выборе библиотеки были рассмотрены следующие параметры: совместимость с ОС, язык программирования, возможность настройки различных нейронных сетей, в том числе сверточных, простота настройки, скорость. Библиотека TensorFlow проста в установке и использовании, возможна настройка нейронных сетей, в том числе сверточных и рекуррентных. TensorFlow выделяется среди остальных библиотек по легкости построения моделей для сверточных нейронных сетей. Используемый язык программирования Python. По скорости TensorFlow не уступает другим библиотекам. Также для данной библиотеки существует большое количество отлично написанных обучающих материалов, их легко понять и использовать на практике.

Также для настройки сети при помощи библиотеки TensorFlow был использован Anaconda — высокопроизводительный дистрибутив Python, содержащий более 100 популярных пакетов Python, R и Scala.

2.2 Проектирование программного обеспечения для определения авторства

Проектируемое программное обеспечение для определения авторства состоит из трех компонентов (пакетов, модулей), показанных на рисунке 17.

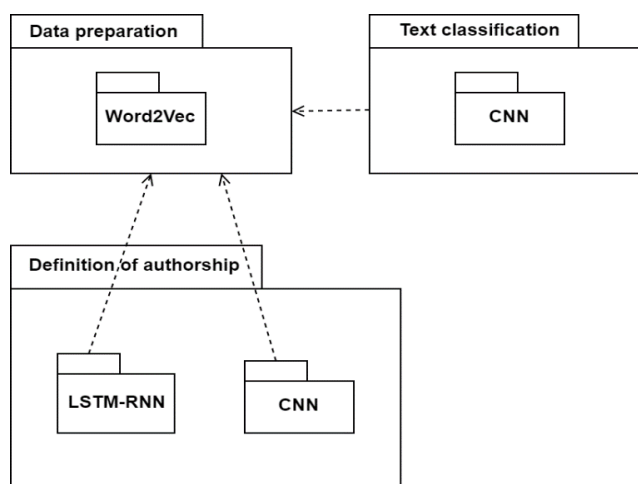


Рисунок 17 – Модульная архитектура программного обеспечения

Как видно из рисунка 17, разрабатываемое программное обеспечение должно состоять из трех модулей.

1) Data preparation представляет собой модуль для считывания исходной выборки из файлов, конфигурирования и обучения модели Word2Vec с настраиваемыми параметрами обучения и записи полученных векторов текстов в файлы;

2) Definition of authorship представляет собой модуль для определения автора текста по фрагменту текста. Модуль состоит из двух подмодулей – CNN и LSTM-RNN, каждый из которых считывает из файла вектора текстов, созданные в модуле Data preparation, конфигурирует, обучает и тестирует свёрточную и рекуррентную нейронную сеть соответственно для определения авторства.

3) Text classification представляет собой модуль для классификации текста на категории (художественная, автобиографическая и научно-популярная литература) на основе фрагмента текста. Модуль считывает из файла вектора текстов, созданные в модуле Data preparation, конфигурирует, обучает и тестирует свёрточную нейронную сеть для классификации текстов.

На рисунке 18 представлена диаграмма вариантов использования разрабатываемого программного обеспечения.

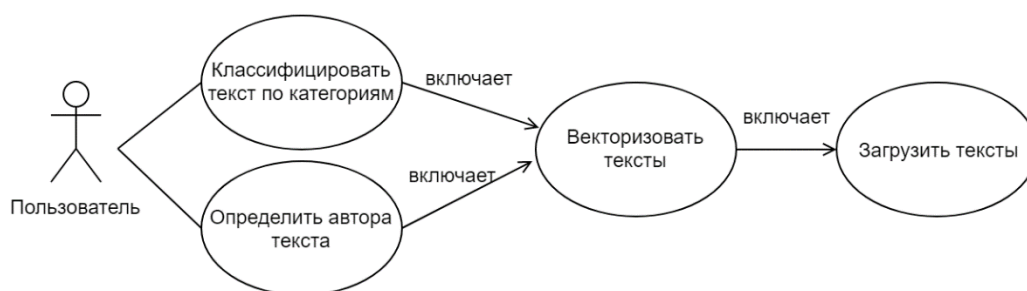


Рисунок 18 – Варианты использования программного обеспечения для определения автора текста

Из рисунка 18 видно, что пользователь должен иметь возможность при помощи разрабатываемого программного обеспечения загружать тексты, векторизовать тексты, обучать модель Word2Vec на основе исходной выборки, классифицировать текст по категориям (художественная,

автобиографическая, научно-художественная литература) и определять автора текста по фрагменту текста.

Обучение модели Word2Vec происходит на основе исходных текстов в модуле подготовки данных (Data preparation). Обучение модели включает векторизацию текстов, которая в свою очередь включает загрузку текстов.

Классификация текстов по категориям (художественная, автобиографическая, научно-художественная литература) должна происходить в модуле и осуществляться сверточной нейронной сетью. Классификация текстов включает векторизацию текстов, которая в свою очередь включает загрузку текстов.

Определение автора текста должно происходить в модуле Definition of authorship в подмодулях LSTM-RNN и CNN, и осуществляться рекуррентной LSTM нейронной сетью и сверточной нейронной сетью соответственно. Определение авторства включает векторизацию текстов, которая в свою очередь включает загрузку текстов.

2.2.1 Реализация сверточной нейронной сети

При реализации сверточной нейронной сети были разработаны и использованы классы, приведенные на рисунке 19.

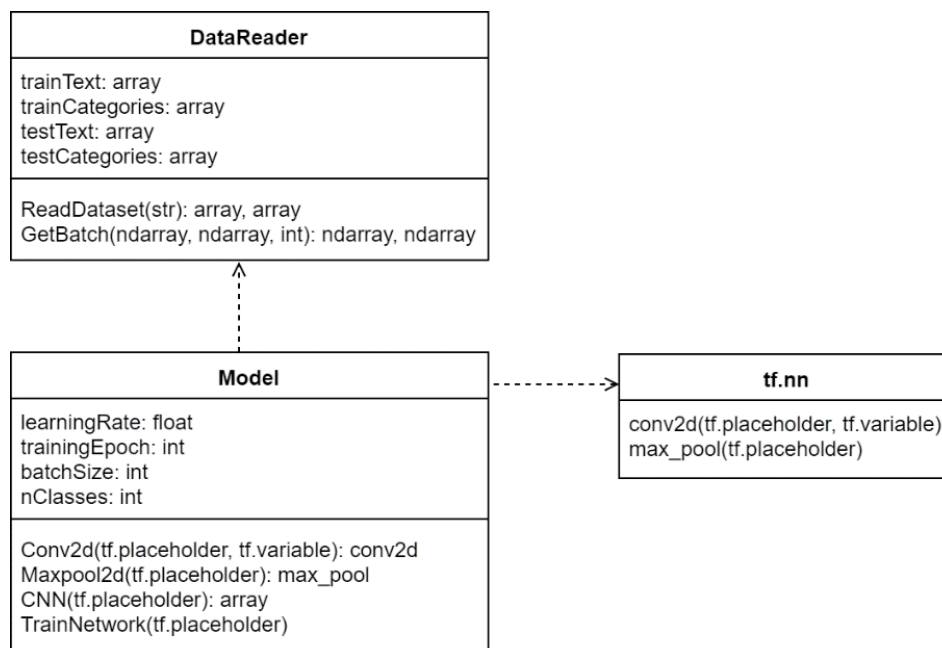


Рисунок 19 – Классы разработанной сверточной нейронной сети

Из рисунка 19 видно, что модуль, реализующий сверточную нейронную сеть, имеет 2 основных класса:

- `DataReader` – класс, в котором определены методы для чтения исходной выборки, такие как `ReadDataset()`, который принимает на вход строку с именем файла, а на выходе возвращает два массива – текстов и категорий. Также в классе определен метод `GetBatch()`, который принимает на вход два массива, полученных в методе `ReadDataset()` и размер блока входной информации, которая будет подаваться на вход сети (`batchSize`).

- `Model` – класс, описывающий модель сверточной нейронной сети. Класс модели реализует методы расчета сверточных слоев нейронной сети – `Conv2d()`, создания субдискретизирующих слоев – `MaxPool2d()`, которые вызывают методы класса `tf.nn` для инициализации слоев сверточной нейронной сети. Метод `CNN()` получает на вход исходную выборку, инициализирует веса и слои нейронной сети в соответствии с конкретной выбранной конфигурацией. Метод `TrainNetwork()` получает на вход исходную выборку и запускает процесс обучения нейронной сети при помощи метода библиотеки `Tensorflow` `tf.train()`.

2.2.2 Проектирование и реализация рекуррентной нейронной сети для определения авторства

2.2.1.1 Архитектура рекуррентной нейронной сети

В качестве рекуррентной нейронной сети была выбрана LSTM-RNN сеть.

Рекуррентная нейронная сеть состоит из 10 ячеек, на вход каждой ячейки подается набор векторов исходных текстов, как показано на рисунке 20.

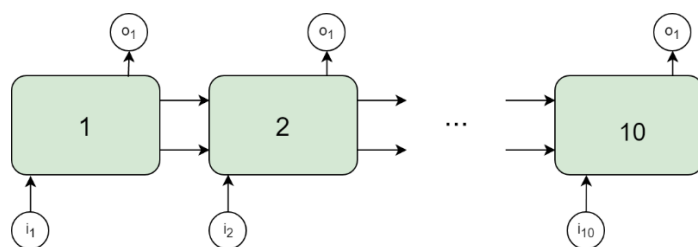


Рисунок 20 – Архитектура реализуемой рекуррентной LSTM нейронной сети

Из рисунка 20 видно, что на вход каждой ячейки LSTM-RNN подается входные данные – $i_1, i_2, \dots, i_{10}$, представляющие собой набор векторов текстов исходной выборки. Каждая ячейка имеет выход – $o_1, o_2, \dots, o_{10}$, представляющий собой предсказание сети относительно того, кто является автором текстов.

2.2.1.2 Реализация рекуррентной нейронной сети

При реализации рекуррентной LSTM нейронной сети были разработаны классы, приведенные на рисунке 21.

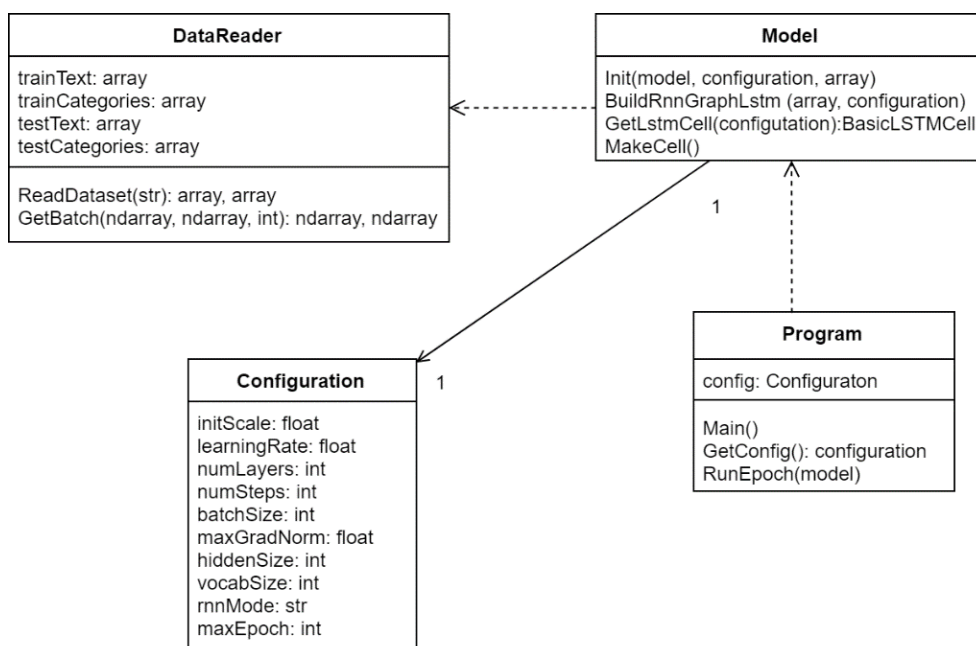


Рисунок 21 – Классы разработанной рекуррентной LSTM нейронной сети

Из рисунка 21 видно, что модуль, реализующий рекуррентную LSTM нейронную сеть, имеет 4 основных класса:

- DataReader – класс, в котором определены методы для чтения исходной выборки, такие как ReadDataset(), который принимает на вход строку с именем файла, а на выходе возвращает два массива – текстов и

категорий (в данном случае категории – это авторы текстов). Также в классе определен метод `GetBatch()`, который принимает на вход два массива, полученных в методе `ReadDataset()` и размер блока входной информации, которая будет подаваться на вход сети (`batchSize`), он заранее определен в конфигурации сети. На выходе метод возвращает готовый блок исходных данных, который далее подавался на вход рекуррентной нейронной сети.

– `Configuration` – класс, в котором определены все необходимые для конфигурации нейронной сети параметры: `initScale` – шкала начальной инициализации весов нейронов рекуррентной нейронной сети, `learningRate` – инициализируемое значение коэффициента обучения сети, `numLayers` – количество слоев LSTM сети, `numSteps` – количество неразвернутых шагов LSTM сети, `maxGradNorm` – предельно допустимая норма градиента, `hiddenSize` – количество LSTM ячеек внутри одного нейрона, `vocabSize` – размер словаря текстов (берется из модели `WordToVec`), `rnnMode` – низкий уровень реализации LSTM ячейки (в Tensorflow предполагается один из вариантов: `CUDNN`, `BASIC` или `BLOCK`, эти уровни представляют различные классы LSTM ячеек), `maxEpoch` – количество эпох обучения. Некоторые из параметров сети задаются один раз, такие как `maxGradNorm`, `vocabSize`, `rnnMode`, `initScale`. Другие же, например, `learningRate` или `maxEpoch` необходимо менять для достижения лучшего результата.

– `Model` – класс, описывающий модель рекуррентной LSTM нейронной сети. Класс модели реализует методы инициализации модели – `Init()`, который создает модель нейронной сети, с соответствующей конфигурацией, инициализирует веса и функцию потерь, задает оптимизатор градиентного спуска. Метод `GetLstmCell()` получает на вход конфигурацию сети и в зависимости от параметра `rnnMode` вызывает различные методы класса `tf.contrib.RNN`. Например, при параметре `rnnMode` равном `BASIC` вызывается метод `BasicLSTMCell()`, который возвращает объект базовой ячейки LSTM сети. Метод `MakeCell()` вызывает метод `GetLstmCell()`. Метод

BuildGraphRnnLSTM() вызывает метод MakeCell() и инициализирует столько ячеек, сколько указано в параметре hiddenLayers.

– Program – класс, описывающий методы Main(), GetConfig(), который возвращает конфигурацию текущей нейросетевой модели и RunEpoch(), который принимает на вход нейросетевую модель и запускает её обучение.

3 Обучение и тестирование

Разрабатываемое программное обеспечение по определению авторства решает две задачи: классификации текстов и определения авторства.

3.1 Задача классификация текстов

Для классификации текстов была настроена и протестирована сверточная нейронная сеть.

3.1.1 Настройка сверточной нейронной сети для классификации текстов

3.1.1.1 Описание исходной выборки для задачи классификации текстов

В качестве исходной выборки были использованы тексты десяти зарубежных авторов.

Выборка содержит тексты различных категорий:

- автобиографическая литература;
- научно-популярная литература;
- художественная литература.

Выборка была сформирована вручную. Текст выборки представлен на английском языке. Тексты имеют различное количество слов, средний размер текста – 10000 слов. На каждого автора приходится по 300 текстов. Выборка была разделена на тренировочную, составляющую 80% от общего объема, и тестовую – 20%. Таким образом, на тренировочную выборку приходится 240 текстов, а на тестовую – 60. Тексты в обоих типах выборок были

перемешаны случайным образом. Обучение проходит в пакетном режиме. Экспериментально было установлено, что наилучшие результаты обучения достигаются при размере пакета 64 входных текста за итерацию.

3.1.1.2 Описание формирования модели Word2Vec и обработка выборки

Так как нейронные сети способны работать только с числовой информацией тексты обучающей выборки необходимо было представить в цифровом виде. Для этого была сформирована модель Word2Vec.

Для создания модели Word2Vec была использована реализация технологии из библиотеки Gensim на языке программирования Python. Параметры обучения представлены ниже:

- Архитектура – Skip-gram. Данный алгоритм работает медленнее, чем CBOW, однако лучше работает для относительно редких слов, что является преимуществом, т.к. уникальных слов в выборке, используемой в работе, большое количество.
- Функция активации – Иерархический софтмакс. Данная функция хорошо подходит для создания модели относительно редких слов. Негативное сэмплирование лучше моделирует более частотные слова.
- Снижение количества часто встречающихся слов: 0.00001.
- Размер векторов слов – 300.
- Контекст (как много слов из контекста обучающий алгоритм должен принимать во внимание) – 10.
- Минимальное количество слов (слова, количество появлений которых меньше заданного игнорируется и не добавляется в словарь) – 1. (Такое количество выбрано для создания максимально подробного описания текста).
- На основе расширенной выборки была получена модель, размер словаря которой составил 9696 слов, где каждое слово является вектором размера 300.

Тексты из выборки имеют различную длину. Необходимо было найти способ взять отдельные вектора слов и преобразовать их в набор характеристик одинаковой длины для всех текстов. Был использован метод усреднения всех векторов слов в каждом тексте. В итоге каждый текст представлял собой вектор размерности 300.

3.1.1.3 Обучение и тестирование сверточной нейронной сети для классификации текста

Сверточная нейронная сеть была реализована на языке Python с использованием библиотеки TensorFlow [13].

При реализации модуля классификации текстов на основе сверточных нейронных сетей была восстановлена нейронная сеть с архитектурой, представленной в таблице 1. В ходе экспериментов она показала наилучший результат [14].

Таблица 1 – Архитектура сверточной нейронной сети для задачи классификации

№ слоя	Количество нейронов	Размер ядра свертки	Размер ядра Max-pooling
1	16	5	2
2	16	5	2
3	256	—	—
4	3	—	—

Как видно из таблицы, архитектура представляет собой чередование 4 слоев. Двух сверточных и двух полносвязных. На последнем слое нейронной сети три нейрона, каждый из которых представляет одну из категорий текстов: автобиографическая, научно-популярная и художественная литература.

На вход нейронной сети на каждой итерации подавалась матрица размера 64×300 . Через каждые 20 итераций производилось тестирование, при котором подавалась вся тестовая выборка, размер матрицы которой 64×300 . Первоначально веса нейронной сети инициализируются случайным образом из нормального распределения. Стандартное отклонение нормального

распределения равно 1. Математическое ожидание распределения равно 0. Функция активации на всех слоях кроме последнего – ReLU, на последнем – Softmax. Оптимизатор градиентного спуска – AdamOptimizer.

Также в ходе экспериментов сверточная нейронная сеть обучалась с различными параметрами. В процессе обучения и тестирования изменялись следующие параметры: коэффициент скорости обучения, количество эпох обучения, размер входной матрицы.

В таблице 2 представлены результаты эксперимента при постоянной скорости обучения 0,001 и входной матрице размером 64×300. В таблице приведены усредненные значения точности по 10 запускам.

Таблица 2 – Параметры и точность нейронной сети при скорости обучения 0,001

Скорость обучения	Количество итераций	Текстов за итерацию	Точность
0,001	10	64	42,01
0,001	20	64	64,13
0,001	50	64	65,67
0,001	100	64	66,18
0,001	120	64	70,10
0,001	150	64	69,09

Из таблицы видно, что точность классификации при скорости обучения 0,001 не превышает 70%. Также можно заметить, что при увеличении количества итераций точность растет, однако наблюдаемый рост низкий и после увеличения до 150 итераций заметно снижение точности.

Исходя из этого было решено увеличить скорость обучения до 0,01. В таблице 3 приведены усредненные значения по 10 запускам.

Таблица 3 – Параметры и точность нейронной сети при скорости обучения 0,01

Скорость обучения	Количество итераций	Текстов за итерацию	Точность
0,01	10	64	70,34
0,01	20	64	74,20
0,01	50	64	80,12
0,01	100	64	90,15
0,01	120	64	75,12

Из таблицы 3 видно, что точность классификации при скорости обучения 0,01 и 100 итерациях точность классификации достигает 90%. Также можно заметить, что при 120 итерациях наблюдается переобучение.

Далее коэффициент обучения был увеличен до 0,1. В таблице 4 приведены усредненные значения по 10 запускам. Так как в предыдущих экспериментах на 10 итерациях не было достигнуто высокой точности в дальнейшем нейронная сеть при данном количество итераций не тестировалась.

Таблица 4 – Параметры и точность нейронной сети при скорости обучения 0,1

Скорость обучения	Количество итераций	Текстов за итерацию	Точность
0,1	20	64	72,40
0,1	50	64	70,13

Из таблицы 4 видно, что точность классификации при скорости обучения 0,1 не превышает точности при коэффициенте обучения 0,01.

При дальнейшем увеличении скорости обучения точность перестала изменяться и её усредненное по 10 запускам значение не превышало 70%.

Также было протестировано влияние размера входной матрицы (в таблицах – количество текстов за итерацию). В таблице 5 приведены усредненные значения по 10 запускам.

Таблица 5 – Параметры и точность нейронной сети при различных размерах входной матрицы

Скорость обучения	Количество итераций	Текстов за итерацию	Точность
0,01	100	8	73,01
0,01	100	16	75,20
0,01	100	32	80,47
0,01	100	64	90,15
0,01	100	128	75,14

Из таблицы 5 видно, что размер входной матрицы также влияет на точность классификации.

Таким образом, после проведенных экспериментов сверточная нейронная сеть была обучена и протестирована на следующих параметрах:

- коэффициент скорости обучения – 0,01;
- количество итераций – 100;
- количество текстов за итерацию – 64.

3.1.1.4 Результаты обучения и тестирования сверточной нейронной сети для классификации текстов

Результат обучения и тестирования сверточной нейронной сети показан на рисунке 22.

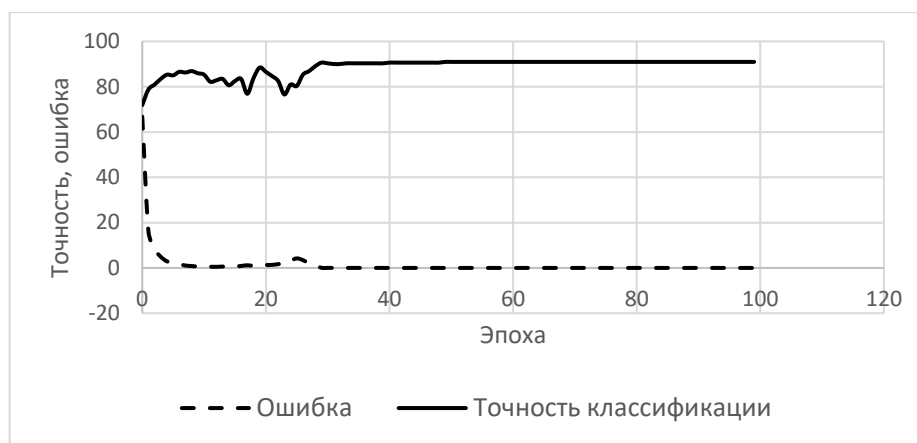


Рисунок 22 – Динамика точности классификации текстов

Сплошная линия характеризует точность классификации, прерывистая – функцию ошибки, которая минимизируется в процессе обучения. Из графика видно, что максимальная точность классификации текста достигла 90%. Таким образом, можно сказать, что сеть успешно справляется с задачей классификации текста.

3.2 Задача определения авторства

Для задачи определения автора по фрагменту текста были использованы сверточные и рекуррентные нейронные сети.

3.2.1 Настройка сверточной нейронной сети для определения автора текста

3.2.1.1 Описание исходной выборки

Изначально в качестве исходной выборки были взяты тексты двух зарубежных авторов новостных рубрик. Текст выборки представлен на английском языке. Тексты имеют различное количество слов, средний размер текста – 1000 слов. На каждого автора приходится по 50 текстов. Выборка была разделена на тренировочную, составляющую 80% от общего объема, и тестовую – 20%. Таким образом, на тренировочную выборку приходится 80 текстов, а на тестовую – 20. Тексты в обоих типах выборок были перемешаны случайным образом.

Для дальнейшей работы было принято решение увеличить обучающую выборку из двух авторов как по количеству авторов (до десяти), количеству и размерам текстов, так и по категориям текстов: художественная, автобиографическая и научно-популярная литература, чтобы в дальнейшем можно было добавить дополнительный признак и производить более точную классификацию. Расширенная выборка содержит тексты 10 зарубежных авторов. Текст выборки представлен на английском языке. Тексты имеют различное количество слов, средний размер текста – 10000 слов. На каждого автора приходится по 300 текстов. Выборка была разделена на тренировочную, составляющую 80% от общего объема, и тестовую – 20%. Таким образом, на тренировочную выборку приходится 240 текстов, а на тестовую – 60. Тексты в обоих типах выборок были перемешаны случайным образом.

3.2.1.2 Описание формирования модели Word2Vec и обработка выборки

Для создания модели Word2Vec была использована реализация технологии из библиотеки Gensim на языке программирования Python. Параметры обучения представлены ниже:

- Архитектура – Skip-gram. Данный алгоритм работает медленнее, чем CBOW, однако лучше работает для относительно редких слов, что является преимуществом, т.к. необходимо выявить слова характерные для стиля того или иного автора.

- Функция активации – Иерархический softmax. Данная функция хорошо подходит для создания модели относительно редких слов. Негативное сэмплирование лучше моделирует более частотные слова.

- Удаление слов с высокой частотой встречаемости (downsampling): частота, выше, чем 0.00001.

- Размер векторов слов – 300.

- Контекст (как много слов из контекста обучающий алгоритм должен принимать во внимание) – 10.

- Минимальное количество слов (слова, количество появлений которых меньше заданного игнорируется и не добавляется в словарь) – 1. (Такое количество выбрано для создания максимально подробного описания текста).

- На основе имеющихся данных была получена модель, размер словаря которой составил 5047 слов, где каждое слово является вектором размера 300.

- На основе расширенной выборки (с десятью авторами) была получена модель, размер словаря которой составил 9696 слов, где каждое слово является вектором размера 300.

Для работы с выборкой из десяти авторов некоторые параметры сети были изменены:

- Удаление слов с высокой частотой встречаемости (downsampling): частота, выше, чем 0.000001. (Такое значение было выбрано для более очевидного различия в векторах разных классов).

- Минимальное количество слов (слова, количество появлений которых меньше заданного игнорируется и не добавляется в словарь) – 10.

(Такое количество выбрано для создания максимально подробного описания текста).

3.2.1.3 Обучение и тестирование сверточной нейронной сети для определения авторства

Сверточная нейронная сеть была реализована на языке Python с использованием библиотеки TensorFlow.

Были протестированы архитектуры сверточной нейронной сети, представленной в таблицах 6 – 8. Обучение сети происходило при одинаковых параметрах: количество итераций – 20, скорость обучения – 0,01. Обучение происходило на выборке из двух авторов.

Таблица 6 – Архитектура сверточной нейронной сети

№ слоя	Количество нейронов	Размер ядра свертки	Размер ядра Max-pooling
1	16	5	2
2	32	—	—
3	2	—	—

Средняя точность сети при 10 запусках составила 75 %.

Далее был добавлен еще один сверточный слой. Из таблицы 7 видно, что это дало небольшой прирост точности.

Таблица 7 – Архитектура сверточной нейронной сети

№ слоя	Количество нейронов	Размер ядра свертки	Размер ядра Max-pooling
1	16	5	2
2	64	5	2
3	64	—	—
4	2	—	—

Средняя точность сети при 10 запусках составила 78 %.

Таблица 8 – Архитектура сверточной нейронной сети

№ слоя	Количество нейронов	Размер ядра свертки	Размер ядра Max-pooling
1	16	5	2
2	16	5	2
3	256	—	—
4	2	—	—

Средняя точность сети при 10 запусках составила 81 %.

В связи с наибольшей точностью, была выбрана архитектура сети, представленная в таблице 8.

В ходе экспериментов сверточная нейронная сеть обучалась с различными параметрами. В таблице 9 представлены наиболее характерные параметры обучения и точности определения автора для выборки с двумя авторами. В таблице представлены усредненные значения точности по 10 запускам.

Таблица 9 – Параметры обучения и точность для выборки из двух авторов

Количество итераций	Текстов за итерацию	Скорость обучения	Точность
20	8	0,001	65,07
	16	0,001	65,51
	32	0,001	71,01
	64	0,001	72,14
50	8	0,01	78,10
	16	0,01	80,09
	32	0,01	80,67
	64	0,01	80,54
100	8	0,01	87,13
	16	0,01	87,70
	32	0,01	87,98
	64	0,01	90,05
120	8	0,01	85,15
	16	0,01	85,17
	32	0,01	88,38
	64	0,01	86,10

Из таблицы 9 видно, что при 20 итерациях точность меняется незначительно. При этом количество пакетов, подаваемых на вход сети на точность незначительно, вследствие недостаточного количества итераций. При увеличении количества итераций до 50 и скорости обучения до 0,01 точность определения авторства увеличилась до 80%. При увеличении количества итераций до 100 точность максимальная точность обучения достигла 90% при 64 текстах за итерацию. Далее количество итераций было увеличено до 120. При этом точность определения авторства остановила свой

рост по сравнению со 100 итерациями. Таким образом, был найден максимум точности.

Таким образом, на выборке из двух авторов, наибольшую точность показала сеть со следующими параметрами обучения. На вход нейронной сети на каждой итерации подавалась матрица размера 64×300 . Всего итераций было произведено 100. Первоначально веса нейронной сети инициализируются случайным образом по закону нормального распределения. Функция активации на всех слоях кроме последнего – ReLU, на последнем – Softmax. Оптимизатор градиентного спуска – AdamOptimizer, скорость обучения – 0,01.

В таблице 10 представлены наиболее характерные параметры обучения и точности определения автора для выборки с десятью авторами. В таблице представлены усредненные значения точности по 10 запускам.

Таблица 10 – Параметры обучения и точность на выборке из десяти авторов

Количество итераций	Текстов за итерацию	Скорость обучения	Точность
100	64	0,01	65,07
100	128	0,01	55,31
100	64	0,001	70,01
100	32	0,001	70,42
120	64	0,001	78,10
150	64	0,001	76,06

Таким образом, на выборке из десяти авторов, наибольшую точность показала сеть со следующими параметрами обучения. На вход нейронной сети на каждой итерации подавалась матрица размера 64×300 . Всего итераций было произведено 120. Первоначально веса нейронной сети инициализируются случайным образом по закону нормального распределения. Функция активации на всех слоях кроме последнего – ReLU, на последнем – Softmax. Оптимизатор градиентного спуска – AdamOptimizer, скорость обучения – 0,001.

3.2.1.4 Результаты обучения и тестирования сверточной нейронной сети для определения авторства

Результат обучения и тестирования сверточной нейронной сети показан на рисунке 23.

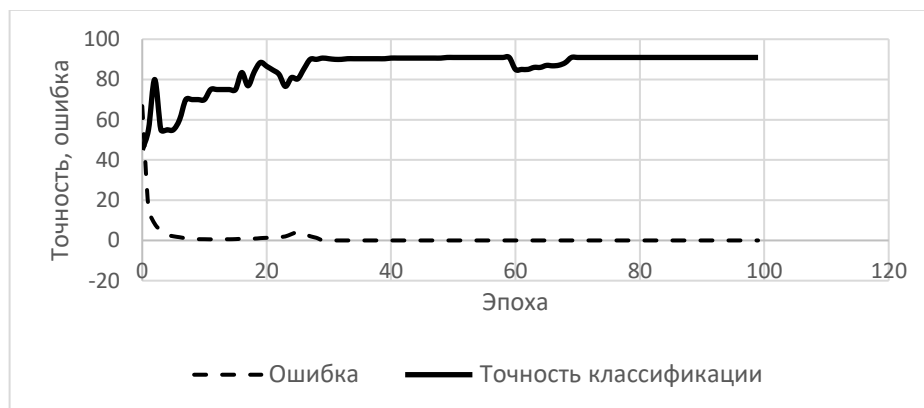


Рисунок 23 – Результат обучения и точность тестирования сети на выборке из двух авторов

Сплошная линия характеризует точность классификации, прерывистая – функцию ошибки, которая минимизируется в процессе обучения. Из графика видно, что максимальная точность классификации текста достигла 90%. Таким образом, можно сказать, что сеть успешно справляется с задачей определения автора по фрагменту текста.

Результат обучения и тестирования сверточной нейронной сети показан на рисунке 24.

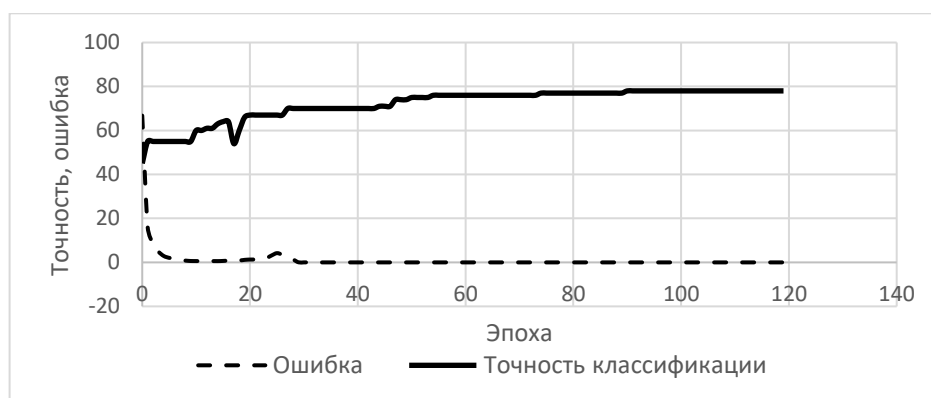


Рисунок 24 – Результат обучения и точность тестирования сети на выборке из десяти авторов

Сплошная линия характеризует точность классификации, прерывистая – функцию ошибки, которая минимизируется в процессе обучения. Из

графика видно, что максимальная точность классификации текста достигла 78%.

3.2.2 Настройка рекуррентной нейронной сети для определения автора текста

3.2.1.1 Описание формирования модели Word2Vec и обработка выборки

Для создания модели Word2Vec также была использована реализация технологии из библиотеки Gensim на языке программирования Python. Параметры обучения представлены ниже:

- Архитектура – Skip-gram. Данный алгоритм работает медленнее, чем CBOW, однако лучше работает для относительно редких слов, что является преимуществом, т.к. необходимо выявить слова характерные для стиля того или иного автора.
- Функция активации – Иерархический softmax. Данная функция хорошо подходит для создания модели относительно редких слов.
- Снижение количества часто встречающихся слов (downsampling): 0.000001. (Такое значение было выбрано для более очевидного различия в векторах разных классов).
- Размер векторов слов – 300.
- Контекст (как много слов из контекста обучающий алгоритм должен принимать во внимание) – 10.
- Минимальное количество слов (слова, количество появлений которых меньше заданного игнорируется и не добавляется в словарь) – 10. (Такое количество выбрано для создания максимально подробного описания текста).

На основе расширенной выборки (с десятью авторами) было получено 10 моделей для того, чтобы на каждой итерации рекуррентной сети подавался новый пакет векторов. Размер словаря каждой модели составил 9696 слов, где каждое слово является вектором размера 300.

3.2.1.2 Обучение и тестирование рекуррентной LSTM нейронной сети

Рекуррентная LSTM нейронная сеть была реализована на языке Python с использованием библиотеки TensorFlow.

В качестве архитектуры ячейки рекуррентной нейронной сети была использована базовая реализация ячейки LSTM сети библиотеки TensorFlow – класс BasicLSTMCell, где в качестве параметра конструктора передается количество блоков ячейки. Базовая реализация представляет собой ячейку, рассмотренную в разделе 1.3.2. Ячейка реализованной сети состоит из 10 блоков, на каждую из которой подавался файл обучающей выборки.

Базовая реализация LSTM сети была протестирована с различными параметрами, такими как коэффициент скорости обучения, количество эпох обучения. Количество слоев и блоков не изменялись и соответствуют базовой реализации. В таблице 11 показаны результаты тестирования, которые показали изменения в точности. В таблице указаны усредненные значения точности по 10 запускам.

Таблица 11 – Параметры обучения и точность

Количество итераций	Скорость обучения	Точность
20	0,01	80,17
20	0,1	83,15
20	0,7	85,38
13	1	90,05

Таким образом, на выборке из десяти авторов, наибольшую точность показала сеть со следующими параметрами обучения. Всего итераций было произведено 13. Функция активации Softmax. Оптимизатор градиентного спуска – GradientDescentOptimizer, скорость обучения – 1. Результаты согласуются с выводами других исследователей [14].

3.2.1.3 Результаты обучения и тестирования рекуррентной LSTM нейронной сети

Результат обучения и тестирования рекуррентной LSTM нейронной сети показан на рисунке 25.

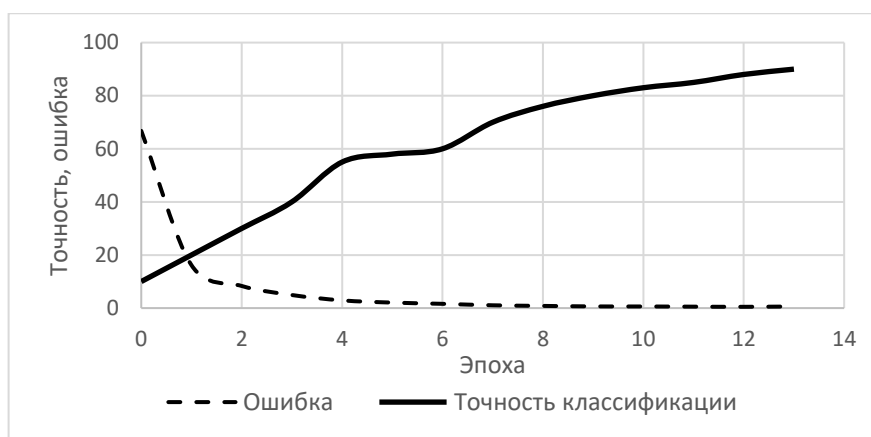


Рисунок 25 – Результат обучения и точность тестирования рекуррентной LSTM нейронной сети

Сплошная линия характеризует точность классификации, прерывистая – функцию ошибки, которая минимизируется в процессе обучения. Из графика видно, что максимальная точность классификации текста достигла 90,3%. Таким образом, можно сказать, что сеть успешно справляется с задачей определения автора по фрагменту текста.

3.2.1.4 Результаты обучения и тестирования рекуррентной LSTM нейронной сети с предварительной классификацией текстов

Для увеличения точности определения авторства была проведена предварительная классификация текстов. После добавлении нового признака в исходную выборку (категории текста: автобиографическая, научно-популярная, художественная литература) нейронная сеть была обучена и протестирована при тех же параметрах и архитектуре, что и в разделе 3.2.1.3.

Результат обучения и тестирования рекуррентной LSTM нейронной сети показан на рисунке 26.

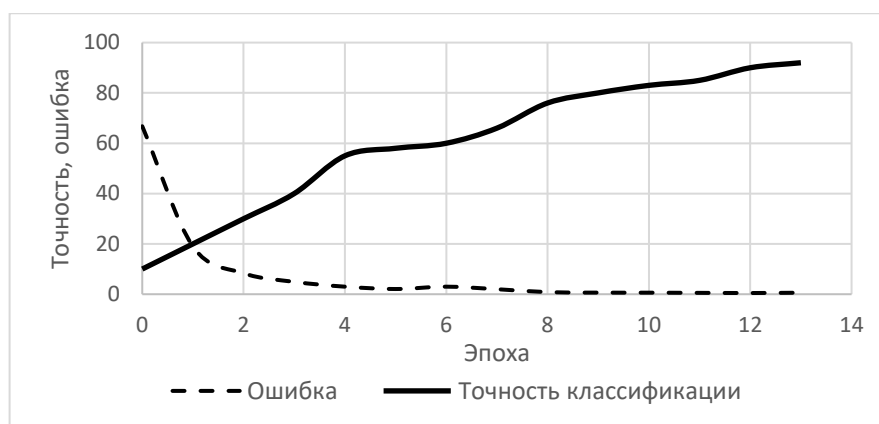


Рисунок 26 – Результат обучения и точность тестирования рекуррентной LSTM нейронной сети с предварительной классификацией текста

Сплошная линия характеризует точность классификации, прерывистая – функцию ошибки, которая минимизируется в процессе обучения. Максимальная точность классификации текста достигла 91%.

3.2.3 Сравнение результатов сверточной и рекуррентной нейронной сети для определения авторства

Исходя из сравнения максимальной точности определения авторства сверточной нейронной сети и рекуррентной LSTM нейронной сети текста можно сказать, что лучший результат показала рекуррентная сеть.

Реализованная рекуррентная LSTM нейронная сеть имеет базовую архитектуру ячейки. Ячейка реализованной сети состоит из 10 блоков, на каждую из которой подавался файл обучающей выборки. Сеть имеет следующие параметры обучения:

- всего итераций – 13;
- скорость обучения – 1.

Первоначально веса нейронной сети инициализируются случайным образом по закону нормального распределения. Функция активации Softmax. Оптимизатор градиентного спуска – GradientDescentOptimizer.

Максимальная точность определения авторства составляет 91%.

4 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

4.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

4.1.1 Потенциальные потребители результатов исследования

Разрабатываемое программное обеспечение направлено на автоматизацию процесса определения авторства.

Так как для определения автора текста зачастую приходится обращаться к экспертам, которые могут идентифицировать автора неизвестного текста или определить принадлежность произведения другому автору при помощи характерных языковых особенностей, стилистических приемов, а задача является востребованной и трудоемкой, появляется потребность в создании формальных методов ее решения. Задача установления авторства текстов (задача атрибуции) встречается в различных областях и представляет интерес для филологов, литературоведов, юристов, криминалистов, историков.

Потенциальные потребители программного обеспечения по определению авторства являются предприятия сферы образования или предприятия, занимающиеся исследованиями в литературоведении.

4.1.2 Анализ конкурентных технических решений

В данном подразделе приведен анализ конкурентных технических решений.

В качестве конкурирующих решений выступают следующие программные продукты:

- система «Лингвоанализатор»;
- система «Атрибутор»;
- система «Авторовед».

Далее для анализа были следующие определены факторы конкурентоспособности:

- повышение производительности труда;
- удобство эксплуатации;
- надежность;
- объем базы;
- функциональность.

Далее для каждого фактора конкурентоспособности была определена его важность (по пятибалльной шкале), после чего была произведена оценка каждого конкурирующего решения (по десятибалльной шкале).

На основе оценочной карты, приведенной в таблице 12, была проанализирована конкурентоспособность разработанного программного обеспечения для определения авторства в сравнении с ближайшими конкурентами.

Таблица 12 – Оценочная карта для сравнения конкурентных технических решений (разработок)

№ п/п	Конкуренты	Факторы конкурентоспособности					Итоговая оценка
		Повышение производительности труда	Удобство эксплуатации	Надежность	Объем базы	Функциональность	
1	Система «Лингвоанализатор»	8 1,52	8 1,144	9 2,142	9 1,71	7 1,666	8,182
2	Система «Атрибутор»	7 1,33	7 1,001	8 1,904	10 1,9	8 1,904	8,039
3	Система «Авторовед»	9 1,71	8 1,144	9 2,142	8 1,52	8 1,904	8,42
4	Данная разработка	8 1,52	9 1,287	9 2,142	8 1,52	10 2,38	8,849
	b_j	4	3	5	4	5	21
	w_j	0,19	0,143	0,238	0,19	0,238	-

Из таблицы 12 видно, что итоговая оценка данной разработки составляет 8,849 и превышает оценки конкурирующих решений. Разработанное программное обеспечение уступает некоторым решениям по факторам повышение производительности труда и объему базы, так как в этих решениях в качестве обучающей выборки использовано больше литературных произведений. Несмотря на это, данная разработка имеет преимущество в удобстве эксплуатации, надежности и функциональности, так как она имеет больший спектр решаемых задач и гибкие алгоритмы работы.

Преимуществом данной разработки перед конкурирующими системами является функциональность программного обеспечения. По оценкам факторов конкурентоспособности был построен многоугольник конкурентоспособности, представленный на рисунке 27.

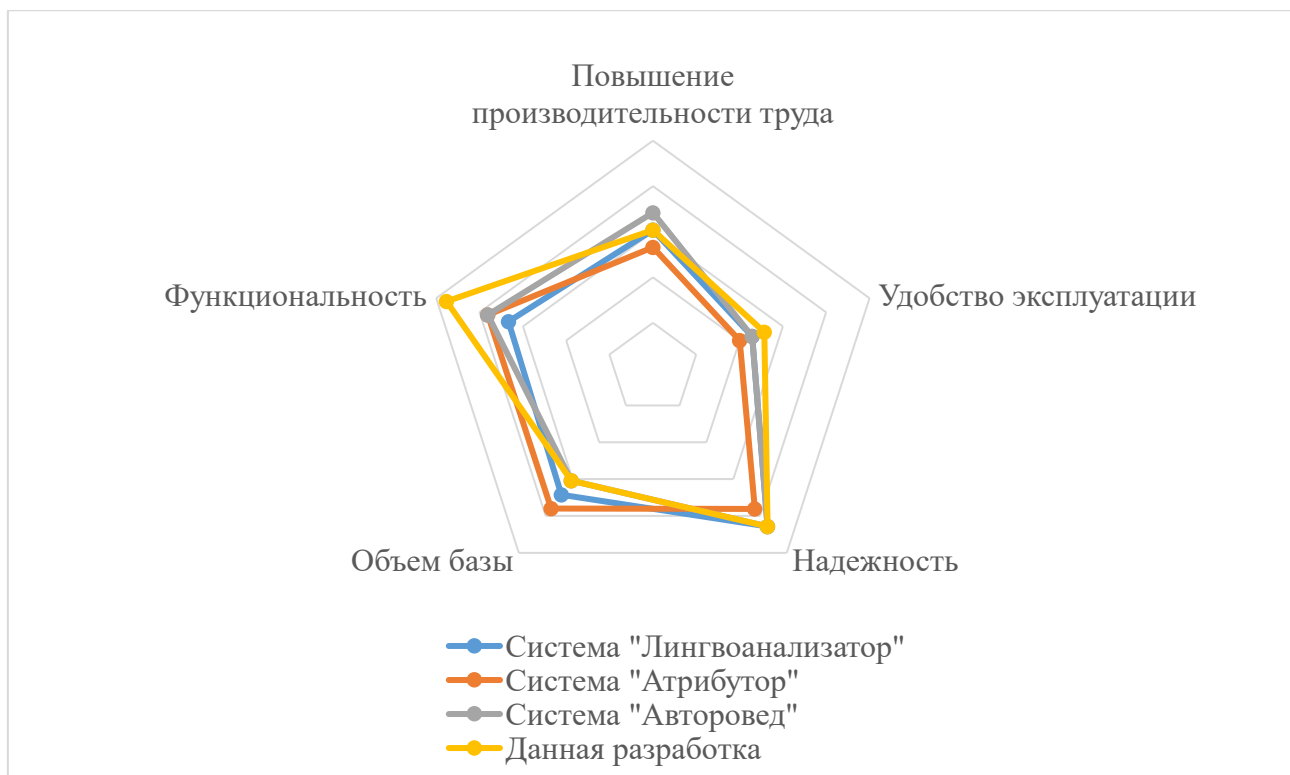


Рисунок 27 – Многоугольник конкурентоспособности

Многоугольник конкурентоспособности наглядно показывает, что разработанное программное обеспечение имеет преимущества в виде повышенной функциональности, удобстве эксплуатации и надежности.

4.1.3 SWOT-анализ

SWOT-анализ представляет собой анализ слабых и сильных сторон, возможностей и угроз исследуемого проекта. Данный анализ применяют для исследования внешней и внутренней среды проекта. SWOT-анализ разработанного программного обеспечения представленный в таблице 13.

Таблица 13 – SWOT-анализ

		Внутренние факторы	
		Сильные стороны проекта:	Слабые стороны проекта:
		1. Актуальность разработки. 2. Низкая стоимость разработки. 3. Конкурентоспособность продукта. 4. Простота интерфейса пользователя.	1. Относительно небольшой охват литературных произведений. 2. Незнание программного обеспечения на рынке.
Внешние факторы	Возможности:	1. Предусмотреть возможность повышения функционала ПО, которая позволит поддерживать конкурентоспособность продукта. 2. Использовать новейшие технологии для обеспечения простоты интерфейса пользователя.	1. Предусмотреть возможность повышения базы литературных произведений. 2. Установить связь с потребителем для продвижения продукта.
	Угрозы:	1. Предусмотреть возможность перехода к новым алгоритмам.	1. Продвигать ПО на рынке. 2. В случае возникновения конкурентов проанализировать причины развития и перенять опыт.

Из SWOT-анализа видно, что основной угрозой проекта является появление более совершенных продуктов, для того, чтобы оставаться конкурентоспособным разработанному программному обеспечению необходимо иметь возможность расширить функционал и изменить алгоритмы работы на более точные (в случае их возникновения).

4.3 Планирование научно-исследовательских работ

4.3.1 Структура работ в рамках научного исследования

Данный раздел содержит описание выполненных работ с указанием исполнителей (таблица 14).

Таблица 14 – Перечень работ и распределение исполнителей

№ работы	Наименование работы	Исполнители работы
1	Выбор научного руководителя бакалаврской работы	Михайлова Е.В.
2	Составление и утверждение темы бакалаврской работы	Иванова Ю.А., Михайлова Е.В.
3	Составление календарного плана-графика выполнения бакалаврской работы	Иванова Ю.А.
4	Подбор и изучение литературы по теме бакалаврской работы	Михайлова Е.В.
5	Анализ предметной области	Иванова Ю.А., Михайлова Е.В.
6	Проектирование программного обеспечения	Иванова Ю.А., Михайлова Е.В.
7	Изучение алгоритмов и моделей искусственных нейронных сетей	Иванова Ю.А., Михайлова Е.В.
8	Реализация программного обеспечения	Иванова Ю.А., Михайлова Е.В.
9	Выполнение численных экспериментов	Иванова Ю.А., Михайлова Е.В.
10	Согласование выполненной работы с научным руководителем	Иванова Ю.А., Михайлова Е.В.
11	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Михайлова Е.В.
12	Подведение итогов, оформление работы	Михайлова Е.В.

4.3.2 Определение трудоемкости выполнения работ

Основной частью стоимости разработки являются трудовые затраты и поэтому определение трудоемкости работ каждого участника исследования становится неотъемлемой частью. С целью определения трудоемкости использованы следующие показатели: ожидаемое значение продолжительности работ, продолжительность каждой работы в днях,

учитывающая параллельность выполнения работ несколькими исполнителями, длительность каждого этапа работ в календарных днях, которая используется для построения диаграммы Ганта, коэффициент календарности.

Согласно производственному календарю (для 6-дневной рабочей недели) в 2019 году 365 календарных дней, 299 рабочих дней, $T_{пр} = 66$ выходных/праздничных дней.

Коэффициент календарности определяется по следующей формуле:

$$T_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}} \quad (1)$$

По формуле 1 был определен коэффициент календарности 2019 года.

$$T_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}} = \frac{365}{365 - 66} = 1,22$$

Для построения календарного плана-графика необходимо рассчитать временные показатели проведения научного исследования. Все расчеты представлены в таблице 15.

Таблица 15 – Временные показатели проведения научного исследования

Наименование работы	Исполнители работы	Трудоемкость работ, чел-дни			Длительность работ, дни	
		tmin	tmax	тож	Тр	Тк
Выбор научного руководителя бакалаврской работы	Михайлова Е.В.	1	3	1,8	2	2
Составление и утверждение темы бакалаврской работы	Михайлова Е.В.	1,8	2,7	2,16	1	1
	Иванова Ю.А.	0,2	0,3	0,24	1	1
Составление календарного плана-графика выполнения бакалаврской работы	Иванова Ю.А.	1	3	1,8	2	2
Подбор и изучение литературы по теме бакалаврской работы	Михайлова Е.В.	7	10	8,2	8	10
Анализ предметной области	Михайлова Е.В.	7	10	8,2	8	10
	Иванова Ю.А.	1	1	1	1	1

Продолжение таблицы 15

Проектирование программного обеспечения	Михайлова Е.В.	8	12	9,6	10	12
	Иванова Ю.А.	1	3	1,8	2	2
Изучение алгоритмов и моделей искусственных нейронных сетей	Михайлова Е.В.	7	15	10,2	10	12
	Иванова Ю.А.	1	1	1	1	1
Реализация программного обеспечения	Михайлова Е.В.	15	30	21	21	26
	Иванова Ю.А.	1	3	1,8	2	2
Выполнение численных экспериментов	Михайлова Е.В.	4	7	5,2	5	6
	Иванова Ю.А.	1	3	1,8	2	2
Согласование выполненной работы с научным руководителем	Михайлова Е.В.	0,9	4,5	2,34	1	1
	Иванова Ю.А.	0,1	0,5	0,26	1	1
Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Михайлова Е.В.	6	8	6,8	7	8
Подведение итогов, оформление работы	Михайлова Е.В.	6	8	6,8	7	8
Итого	Михайлова Е.В.	63,7	110,4	82,3	80	98
	Иванова Ю.А.	7,3	15,8	10,7	12	14

Исходя из расчетов, видно, что общее количество рабочих дней над проектом студентом – 98 дней, научным руководителем – 14 дней.

4.3.3 Разработка графика проведения научного исследования

Для наглядного представления распределения работ участников проекта и затраченного времени была построена диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ. Построенная диаграмма Ганта представлена на рисунке 28.

Дата начала работы над проектом – 21.01.2019, дата окончания – 02.06.2019. Таким образом, общая длительность работы над проектом составляет 98 календарных дней.

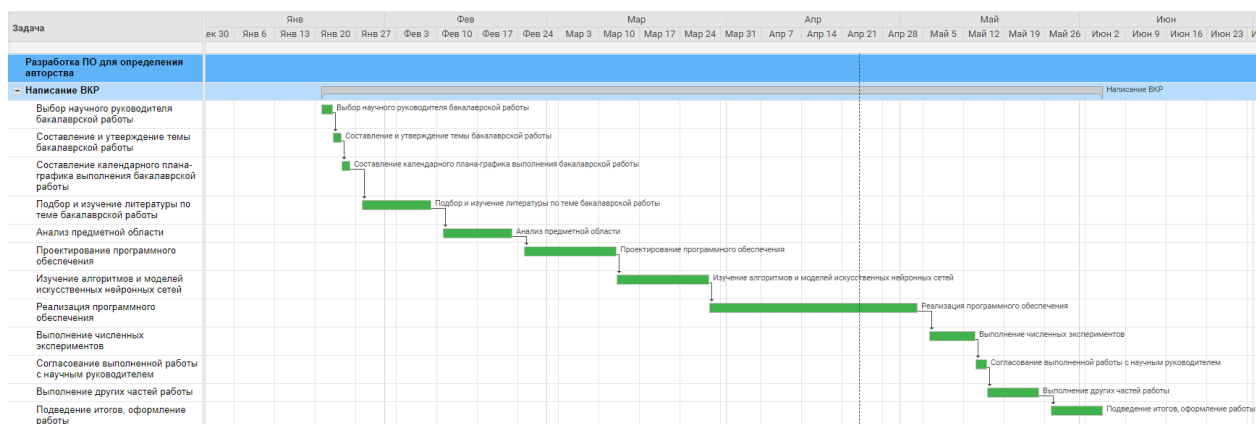


Рисунок 28 – Диаграмма Ганта

4.3.4 Бюджет научно-технического исследования

4.3.4.1 Расчет материальных затрат научно-технического исследования

Данная статья затрат включает в себя затраты на приобретение сырья, материалов, полуфабрикатов и комплектующих со стороны. Также в эту статью включаются транспортные расходы, равные 15% от общей стоимости материальных затрат.

В стоимость материальных затрат могут быть включены затраты на канцелярские принадлежности.

Сумма материальных затрат, необходимых для проведения данного исследования, включает только затраты на канцелярию и равна 2000 руб.

4.3.4.2 Расчет затрат на специальное оборудование для научных (экспериментальных) целей

Затраты на специальное оборудование включают в себя затраты на покупку дорогостоящего оборудования и ПО или же амортизация, если оборудование или ПО есть у студента.

Для оценки затрат на оборудование была рассчитана амортизация использованного оборудования – ПК. Для написания ВКР был использован

ПК стоимостью 60000 рублей. Компьютеры относятся к имуществу со сроком полезного использования от 2 до 3 лет. При расчете использовался максимальный срок – 3 года. Планируемая длительность написания ВКР – 6 месяцев. Амортизация рассчитывается линейным способом.

Норма амортизации рассчитывается по следующей формуле:

$$A_n = \frac{1}{n} * 100\%, \quad (2)$$

где n – срок полезного использования

Годовые амортизационные отчисления рассчитываются по формуле ниже:

$$A_g = C \times A_n, \quad (3)$$

где C – стоимость оборудования

Ежемесячные амортизационные отчисления рассчитываются по формуле ниже:

$$A_m = \frac{A_g}{12} \quad (4)$$

Итоговая сумма амортизации основных средств рассчитывается по следующей формуле:

$$A = A_m \times t, \quad (5)$$

где t – планируемая длительность написания ВКР.

Используя формулы 2 – 5 была рассчитана амортизация на использованный ПК:

- норма амортизации:

$$A_n = \frac{1}{n} * 100\% = \frac{1}{3} \times 100\% = 33,33\%$$

- годовые амортизационные отчисления:

$$A_g = 60000 \times 0,33 = 19800 \text{ рублей}$$

- ежемесячные амортизационные отчисления:

$$A_m = \frac{19800}{12} = 1650 \text{ рублей}$$

- итоговая сумма амортизации основных средств:

$$A = 1650 \times 6 = 9900 \text{ рублей}$$

В таблице 16 приведены затраты на амортизацию.

Таблица 16 – Расчет затрат на амортизацию

Наименование	Затраты, руб.
Амортизация ПК	9900

Таким образом, итоговые затраты на амортизацию составляют 9900 рублей.

4.3.4.3 Основная заработная плата исполнителей темы

В этом подразделе рассчитана заработная плата исполнителей проекта. Данная статья затрат включает в себя заработную плату специалистов, непосредственно участвующих в выполнении проекта.

Для расчета основной заработной платы студента использовался оклад, равный окладу ассистента без степени – 21760 руб. Оклад научного руководителя – 33664 рублей.

При расчете заработной платы необходимо знать баланс рабочего времени, который приведен в таблице 17.

Таблица 17 – Баланс рабочего времени (для 6-дневной недели)

Показатели рабочего времени	Дни
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Заработная плата состоит из основной и дополнительной.

Основная заработная плата рассчитывается по формуле ниже.

$$Z_{осн} = Z_{дн} \times T_r \times (1 + K_{пр} + K_{д}) \times K_r, \text{ где} \quad (6)$$

$Z_{дн}$ – среднедневная заработная плата, руб.

$K_{пр}$ – премиальный коэффициент (0,3);

$K_{д}$ – коэффициент доплат и надбавок (0,2);

K_r – районный коэффициент (для Томска 1,3);

T_r – продолжительность работ, выполняемых работником, раб. Дни

Средняя заработная плата рассчитывается по следующей формуле:

$$З_{дн} = \frac{З_{м \times М}}{F_{д}}, \text{ где} \quad (7)$$

$З_{м}$ – месячный оклад работника, руб.

$М$ – количество месяцев работы без отпуска в течение года:

при отпуске в 48 раб. дней $М=10,4$ месяца, 6-дневная неделя;

$F_{д}$ – действительный годовой фонд рабочего времени персонала, раб. дн.

При помощи формул 6 – 7 были проведены расчеты основной заработной платы студента и научного руководителя проекта.

Среднедневная заработная плата для ассистента составляет:

$$З_{дн} = \frac{З_{м} \times М}{F_{д}} = \frac{21760 \times 10,4}{243} = 931,29 \text{ руб.}$$

Основная заработная плата для ассистента составляет:

$$\begin{aligned} З_{осн} &= З_{дн} \times Тр \times (1 + К_{пр} + К_{д}) \times К_{р} \\ &= 931,29 \times 80 \times (1 + 0,3 + 0,2) \times 1,3 = 145\,281,24 \end{aligned}$$

Среднедневная заработная плата для доцента составляет:

$$З_{дн} = \frac{З_{м} \times М}{F_{д}} = \frac{33664 \times 10,4}{243} = 1440,76 \text{ руб.}$$

Основная заработная плата для доцента составляет:

$$\begin{aligned} З_{осн} &= З_{дн} \times Тр \times (1 + К_{пр} + К_{д}) \times К_{р} \\ &= 1440,76 \times 12 \times (1 + 0,3 + 0,2) \times 1,3 = 33\,713,78 \end{aligned}$$

В таблице 18 приведен расчет основной заработной платы студента и научного руководителя проекта.

Таблица 18 – Расчет основной заработной платы

Исполнители	Здн, руб.	Кпр	Кд	Кр	Тр	Зосн, руб.
Студент	931,29	0,3	0,2	1,3	80	145 281,24
Научный руководитель	1440,76	0,3	0,2	1,3	12	33 713,78
Итого:						178 995

Таким образом, общая сумма основной заработной платы составила 178 995 рублей.

4.3.4.4 Дополнительная заработная плата исполнителей темы

В данном подразделе приведены дополнительные заработные платы студента и научного руководителя проекта. Данная статья включает в себя величину доплат за отклонение от нормальных условий труда, а также выплат, связанных с обеспечением гарантий и компенсаций (например, при совмещении работы с обучением, при предоставлении ежегодного оплачиваемого отпуска).

Дополнительная заработная плата составляет 12–15 % от основной заработной платы. Для расчетов использовалось значение 15%.

Дополнительная заработная плата рассчитывалась по следующей формуле:

$$З_{\text{доп}} = З_{\text{осн}} \times 0,15 \quad (8)$$

Дополнительная заработная плата студента, рассчитанная по формуле 8:

$$З_{\text{доп}} = З_{\text{осн}} \times 0,15 = 145\,281,24 \times 0,15 = 21\,792,18 \text{ руб.}$$

Дополнительная заработная плата научного руководителя, рассчитанная по формуле 8:

$$З_{\text{доп}} = З_{\text{осн}} \times 0,15 = 33\,713,78 \times 0,15 = 5\,057 \text{ руб.}$$

Итоговая дополнительная заработная плата составляет 26 849,18 руб.

4.3.4.5 Отчисления во внебюджетные фонды (страховые отчисления)

В данном подразделе приведены страховые отчисления студента и научного руководителя проекта. Данная статья затрат включает в себя обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования, пенсионного фонда и медицинского страхования от затрат на оплату труда работников.

Страховые отчисления составляют 30% от суммы основной и дополнительной заработной платы.

Страховые отчисления рассчитывались по следующей формуле:

$$Рст = (Зосн + Здоп) \times 0,3 \quad (9)$$

Страховые отчисления студента, рассчитанные по формуле 9:

$$Рст = (Зосн + Здоп) \times 0,3 = (145\,281,24 + 21\,792,18) \times 0,3 = 50122 \text{ руб.}$$

Страховые отчисления научного руководителя, рассчитанные по формуле 9:

$$Рст = (Зосн + Здоп) \times 0,3 = (33\,713,78 + 5\,057) \times 0,3 = 11631,23 \text{ руб.}$$

Итоговая сумма страховых отчислений составляет 61753,23 руб.

4.3.4.8 Накладные расходы

В данном подразделе приведены накладные расходы студента. Данная статья расходов включает в себя прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые и телеграфные расходы, размножение материалов и другие.

Накладные расходы составляют 16% от суммы материальные затрат, затрат на специальное оборудование, затрат на основную заработную плату, затрат на дополнительную заработную плату и страховые взносы.

Накладные расходы рассчитывались по следующей формуле:

$$Рн = (Рм + Рсп + Зосн + Здоп + Рст) \times 0,16, \text{ где} \quad (10)$$

Рм – материальные затраты,

Рсп – затраты на специальное оборудование.

Накладные расходы, рассчитанные по формуле 10:

$$Рн = (Рм + Рсп + Зосн + Здоп + Рст) \times 0,16 = (2000 + 9900 + 178\,995 + 26\,849,18 + 61753,23) \times 0,16 = 44\,719,58 \text{ руб.}$$

Накладные расходы проекта составляют 44 719,58 рублей.

4.3.4.9 Формирование бюджета затрат научно-исследовательского проекта

В данном подразделе приведен расчет бюджета затрат научно-исследовательского проекта (таблица 19). Бюджет состоит из затрат, рассчитанных в предыдущих подразделах (2.4.1 – 2.4.6).

Таблица 19 – Расчет бюджета затрат НТИ

Наименование	Сумма, руб.	Удельный вес, %
Материальные затраты	2000	0,62
Затраты на специальное оборудование	9900	3,1
Затраты на основную заработную плату	178 995	55,2
Затраты на дополнительную заработную плату	26 849,18	8,28
Страховые взносы	61 753,23	19
Накладные расходы	44 719,58	13,8
Общий бюджет	324 216,99	100%

Таким образом, общий бюджет проекта составляет 324 216,99 рублей, из которых 55,2% составляют затраты на основную заработную плату.

4.4 Определение потенциального эффекта исследования

В качестве потенциальных потребителей программного обеспечения можно назвать предприятия сферы образования или предприятия, занимающиеся исследованиями в литературоведении.

Положительный эффект данного исследования состоит в том, что оно автоматизирует процесс определения авторства текста (решает задачу атрибуции текста) и классификации текста. Таким образом, разработка позволит повысить производительность труда экспертов, занимающихся этими задачами и снизить риск ошибки, обусловленный человеческим фактором.

Исходя из проведенного анализа конкурентоспособности можно сделать вывод, что разработка является конкурентоспособной, в

особенности, с точки зрения функциональности и надежности. Уровень конкурентоспособности составляет 8,849.

Общая длительность исследования для студента составляет 80 рабочих дней, что соответствует 98 календарным дням. Общая длительность исследования для научного руководителя составляет 12 рабочих дней, что соответствует 14 календарным дням. Потенциальная стоимость исследования, состоящая из затрат на материальные расходы, затрат на специальное оборудование, основной и дополнительной заработной платы, страховых взносов и накладных расходов составляет 324 216,99 рублей, из которых более 50% составляют затраты на основную заработную плату.

5 Социальная ответственность

Создаваемым программным продуктом является программное обеспечение для определения автора по фрагменту произведения на основе искусственных нейронных сетей. Основной задачей проекта является автоматизация автороведческой экспертизы и, как следствие, облегчение работы специалистов, занимающихся этим процессом. Потенциальные потребители программного обеспечения по определению авторства являются предприятия сферы образования или предприятия, занимающиеся исследованиями в литературоведении.

5.1 Правовые и организационные вопросы обеспечения безопасности

5.1.1 Специальные правовые нормы трудового законодательства

Функции государственного надзора и контроля в организациях независимо от организационно-правовых форм и форм собственности осуществляются специально уполномоченными на то государственными органами и инспекциями согласно федеральным законам.

Трудовой Кодекс Российской Федерации определяет ежедневный и междусменный отдых – это отдых после окончания рабочего дня или смены.

Трудовой Кодекс также определяет продолжительность рабочего дня. Согласно 91 статье ТК РФ нормальная продолжительность рабочего времени не может превышать 40 часов в неделю. Работодатель обязан вести учет времени, фактически отработанного каждым работником.

Каждый работник имеет право на выходные дни, то есть периоды еженедельного непрерывного отдыха. Согласно 110 статье ТК РФ продолжительность еженедельного непрерывного отдыха не может быть менее 42 часов.

Трудовой кодекс РФ не определяет работу за компьютером как вредные условия труда.

Нормативными документами, регламентирующими деятельность работника с использования ПЭВМ, являются СанПиН 2.2.2/2.4.1340-03.

Гигиенические требования к ПЭВМ и организации труда и Инструкция по охране труда при работе с ПК. Нормативным документом СанПин 2.2.2/2.4.1340-03 установлено, что при работе с компьютером, пользователь обязан делать перерыв в размере 15 минут каждый час непрерывной работы [16].

Во время регламентированных перерывов целесообразно выполнять комплексы упражнений и осуществлять проветривание помещения. Не рекомендуется работать за компьютером более 6 часов за смену. Для того чтобы ПЭВМ соответствовали нормам, осуществляется производственный контроль и надзор внутри предприятия-производителя. Эксплуатирующие предприятия также следят за характеристиками используемой аппаратуры.

5.1.2 Организационные мероприятия при компоновке рабочей зоны

Большое значение для профилактики статических физических перегрузок имеет правильная организация рабочего места человека, работающего с ПЭВМ. Рабочее место должно быть организовано в соответствии с требованиями стандартов, технических условий и (или) методических указаний по безопасности труда.

Оно должно удовлетворять следующим требованиям:

- обеспечивать возможность удобного выполнения работ;
- учитывать физическую тяжесть работ;
- учитывать размеры рабочей зоны и необходимость передвижения в ней работающего;
- учитывать технологические особенности процесса выполнения работ.

Невыполнение требований к расположению и компоновке рабочего места может привести к получению пользователем производственной травмы или развития у него профессионального заболевания.

Рабочее место программиста должно соответствовать требованиям СанПин 2.2.2/2.4.1340-03:

– Конструкция оборудования и рабочего места при выполнении работ в положении сидя должна обеспечивать оптимальное положение работающего, которое достигается регулированием высоты рабочей поверхности, высоты сидения, оборудованием пространства для размещения ног и высотой подставки для ног.

– Схемы размещения рабочих мест с персональными компьютерами должны учитывать расстояния между рабочими столами с мониторами: расстояние между боковыми поверхностями мониторов не менее 1,2 м, а расстояние между экраном монитора и тыльной частью другого монитора не менее 2,0 м. Клавиатура должна располагаться на поверхности стола на расстоянии 100-300 мм от края, обращенного к пользователю. Быстрое и точное считывание информации обеспечивается при расположении плоскости экрана ниже уровня глаз пользователя, предпочтительно перпендикулярно к нормальной линии взгляда (нормальная линия взгляда 15 градусов вниз от горизонтали).

– Рабочие места с ПЭВМ при выполнении творческой работы, требующей значительного умственного напряжения или высокой концентрации внимания, рекомендуется изолировать друг от друга перегородками высотой 1,5 – 2,0 м.

5.2 Производственная безопасность

5.2.1 Анализ вредных и опасных факторов, которые может создать объект исследования

В процессе анализа вредных и опасных факторов, были выявлены:

- повышенный уровень шума на рабочем месте;
- повышенный уровень электромагнитных излучений;
- перенапряжение зрительного анализатора.

5.2.1.1 Повышенный уровень шума

Шум — это совокупность звуков, неблагоприятно воздействующих на организм человека и мешающих его работе и отдыху.

К основным источникам шума, при использовании программного обеспечения можно отнести звук работающей системы охлаждения ПЭВМ. В соответствии с нормативным документом СН 2.2.4/2.1.8.562-96 «Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки. Санитарные нормы» установлен допустимый предел уровня шума, не превышающий 50дБА [17].

Воздействие шума отрицательно влияет на работу человека, поскольку вызывает головную боль, быструю утомляемость, ослабляет внимание и ухудшает слух человека.

5.2.1.2 Повышенный уровень электромагнитных излучений

Также персональный компьютер является основным источником повышенных электромагнитных излучений, так как они усиливают риск возникновения сердечно-сосудистых заболеваний, а также негативно влияют на нервную систему.

Оценка опасности воздействия электромагнитного поля на человека производится по величине электромагнитной энергии, поглощаемой телом человека, с учетом электрической и магнитной напряженностей поля.

Практически, при обслуживании даже мощных, электроустановок высокого напряжения магнитная напряженность значительно меньше опасной (в 8 раз), поэтому оценку потенциальной опасности воздействия электромагнитного поля достаточно производить по величине электрической напряженности поля. В соответствии с ГОСТ 12.1.002—84, ССБТ «Электромагнитные поля токов промышленной частоты. Общие требования безопасности», нормы допустимых уровней напряженности электромагнитных полей зависят от времени пребывания человека в контролируемой зоне [18].

Присутствие персонала на рабочем месте в течение 8 ч допускается при напряженности, не превышающей 5 кВ/м. Основными видами средств коллективной защиты от воздействия электромагнитного поля токов промышленной частоты являются стационарные или переносные

заземленные экранирующие устройства. Так как электромагнитное излучение в месте работы не превышает 5 кВ/м, применение экранирующих устройств не требуется.

5.2.1.3 Перенапряжение зрительного анализатора

Перенапряжение зрительного анализатора может возникнуть вследствие длительной работы за компьютером.

Согласно требованиям, приведенным в ГОСТ Р 55710-2013 для рабочих мест, оборудованных персональными компьютерами или мониторами, допустимые значения габаритной яркости осветительного прибора, отражающихся в экранах мониторов, в зависимости от яркости экранов/мониторов приведены в таблице 29 [19].

Таблица 29 – Допустимые значения габаритной яркости ОП

Класс представления информации	Габаритная яркость ОП, кд/м	Яркость экранов/мониторов, кд/м
А - темные знаки на светлом фоне	До 1500 включ. " 3000 "	До 200 включ. Св. 200
В - светлые знаки на темном фоне	" 1000 " " 1500 "	До 200 включ. Св. 200
Примечание - Значения габаритной яркости ОП определяют под углом не менее 65° от вертикали для рабочих мест, в которых находятся дисплейные экраны с углом наклона 150°.		

Согласно СанПиН 2.2.2/2.4.1340-03 освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300-500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк [16].

5.2.2 Анализ вредных и опасных факторов, которые могут возникнуть при производстве объекта на предприятии

При анализе вредных и опасных факторов, которые могут возникнуть при разработке программного обеспечения были выявлены следующие факторы:

- отклонение от показателей микроклимата;
- недостаточная освещенность рабочей зоны.

5.2.2.1 Отклонение от показателей микроклимата в помещении

Отклонение показателей микроклимата может привести к ухудшению самочувствия человека и/или потере концентрации.

Показателями, характеризующими микроклимат в производственных помещениях, являются: температура воздуха; температура поверхностей; относительная влажность воздуха; скорость движения воздуха; интенсивность теплового облучения.

Согласно СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений» работа на компьютере относится к категории Ia – работы с интенсивностью энергозатрат до 120 ккал/ч (до 139 Вт), производимые сидя и сопровождающиеся незначительным физическим напряжением [20].

В производственных помещениях, в которых работа на компьютере является основной, обеспечиваются оптимальные параметры микроклимата (Таблица 30 – Оптимальные величины показателей микроклимата для категории работ Ia (по СанПиН 2.2.4.548-96)).

Таблица 30 – Оптимальные величины показателей микроклимата для категории работ Ia (по СанПиН 2.2.4.548-96).

Период года	Температура воздуха, °С	Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	22-24	21-25	40-60	Не более 0,1
Теплый	23-25	22-26	40-60	Не более 0,1

5.2.2.2 Недостаточная освещенность рабочей зоны

Под освещенностью понимается отношение светового потока, падающего на элемент поверхности, к площади этого элемента. Обозначается освещенность буквой E , измеряется в люксах. Согласно ГОСТ Р 55710-2013 «Освещение рабочих мест внутри зданий. Нормы и методы измерений» средняя освещенность на рабочих местах с постоянным пребыванием людей должна быть не менее 200 лк [19].

Недостаточная освещенность рабочего места увеличивает напряжение глаз сотрудников, что может привести к ухудшению зрения. При плохой освещенности труднее различать цвета, возможно снижение способности к концентрации. Также недостаток света в помещении приводит к снижению уровня работоспособности, бодрости и ухудшает настроение сотрудников. Кроме того, низкое или чересчур пульсирующее освещение может способствовать появлению головных болей или мигреней. К пульсации приводит, как правило, использование газоразрядных ламп, работающих на частоте 50 Гц.

Для комфортной работы сотрудника необходимо отсутствие пульсации света, обеспечение достаточной контрастности в цветопередаче монитора, отсутствие бликов на поверхностях офисного оборудования, а также соответствующее направление светового потока и его спектр.

5.3 Мероприятия по защите персонала предприятия от действия опасных и вредных факторов

5.3.1 Рекомендации по минимизации влияния шума

Для снижения воздействия шума может быть применен метод удаления оператора от объекта удаления. Также для уменьшения шумов, генерируемых персональными компьютерами, рекомендуется проводить их регулярную диагностику, выполнять чистку и устранение неисправностей.

5.3.2 Рекомендации по минимизации влияния электромагнитных излучений

Для защиты пользователя компьютера от данного вида негативного излучения необходимо обеспечить оптимальное расстояние между монитором и работником, равное минимум 500 мм, установленное ГОСТ Р 50948-2001 «Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности» [21].

5.3.3 Рекомендации по минимизации влияния на зрительный анализатор

Для предотвращения негативных последствий длительной работы за компьютером необходимо соблюдать режим труда и отдыха, использовать зрительную гимнастику, обеспечивать безопасное расстояние между экраном монитора и пользователем равным 500-700 мм в соответствии с нормативным документом ГОСТ Р 50948-2001.

5.3.4 Рекомендации по минимизации влияния отклонений от показателей микроклимата в помещении

К мероприятиям по оздоровлению воздушной среды на предприятии относятся организация вентиляции (как естественным, так и механическим путем) и кондиционирования воздуха, отопление комнаты (в зимнее время года). Объем помещений с ЭВМ не должен быть меньше 20 м³/человека.

5.3.5 Рекомендации по минимизации влияния недостаточной освещенность рабочей зоны

Для обеспечения требуемого уровня освещения в помещении необходимо использовать лампы дневного освещения, равномерно распределенные по рабочей зоне.

5.4 Экологическая безопасность

5.4.1 Анализ влияния объекта исследования на окружающую среду

Разработка программного обеспечения не оказывает влияние на гидросферу и атмосферу. Однако, в случае выхода из строя частей или всего персонального компьютера, на котором ведется разработка или осуществляется выполнение ПО, происходит воздействие на литосферу.

Компьютеры необходимо утилизировать в специально отведенных местах, так как они из разных деталей, которые могут негативно сказаться на экологии территории, поэтому выбрасывание их на свалку может привести к серьезным последствиям. К тому же, в компьютерной технике содержится определенная часть драгоценных металлов (палладий, серебро, платина, золото) и веществ, которые можно вторично использовать на производстве (переработка компьютерного пластика, железа, стекла).

Компьютерное оборудование не принадлежит к одному классу опасности, ведь оно содержит детали, которые по-разному будут взаимодействовать с окружающей средой. В компьютерах могут одновременно содержаться и опасные, и безопасные элементы:

- ртутные лампы, используемые в ПК, ноутбуках, мониторах, являются чрезвычайно опасными, поэтому отнесены к I классу;
- платы и аккумуляторы, которые содержат свинец, кадмий или олово, относятся ко II классу опасности;
- трансформаторы и провода – к III классу;
- металлические детали практически безопасны, и им присвоена V степень опасности.

5.4.2 Мероприятия по защите окружающей среды

Согласно ГОСТ Р 55102-2012 сбор различных видов отработавшего оборудования целесообразно осуществлять в отдельные упаковки, способные обеспечить неизменность свойств для дальнейшего хранения и

транспортирования. Таким образом, в отдельные упаковки собирают системные блоки компьютеров, клавиатуры, звукопроигрывающую аппаратуру и радиоприемники [22].

Устаревшее оборудование и неисправная оргтехника по роду своего происхождения относится к классу отходы производства. В свою очередь электронные промышленные отходы подразделяется на два подкласса: вторичное сырье и безвозвратные потери. К вторичному сырью относятся переработка цветных металлов, полимерных материалов, компоненты компьютера, которые после переработки становятся деталями для изготовления новых технических устройств.

Утилизация металлов осуществляется по средствам приемных пунктов, которые реализуют их первичную переработку, следующий этап утилизации решается путем различных методов сепарации. Метод вторичной переработки актуален для полимерных отходов, поскольку утилизация пластика является проблемой глобального характера. Данный метод позволяет сократить количество пластиковых отходов на полигонах, что является эффективным решением существующей экологической проблемы.

Часть отходов, возникших в процессе утилизации оргтехники, являются безвозвратными, поскольку в состав технического оборудования входят элементы, переработка которых на сегодняшний день нецелесообразна. Следовательно, отходы такого вида подвергаются захоронению на специализированных полигонах, что несет существенный урон литосфере нашей планеты.

5.5 Безопасность в чрезвычайных ситуациях

5.5.1 Анализ вероятных ЧС, которые может инициировать объект исследований

В ходе эксплуатации данной системы на рабочем месте могут возникнуть пожары и взрывы, относящиеся к чрезвычайным ситуациям.

Наиболее вероятными чрезвычайными ситуациями являются пожары, которые могут быть вызваны коротким замыканием, перегрузкой,

нагреванием участков электросети, неисправностью оборудования и несоблюдения правил пожарной безопасности.

Пожар может нанести не только вред здоровью, но и материальный ущерб. Применимо к выполняемой работе в случае пожара могут быть уничтожены бумажные документы и\или электронные носители информации. Для защиты информации рекомендуется использовать облачные хранилища данных для данных и документов. Для исходных кодов программ рекомендуется использовать системы контроля версий.

5.5.2 Мероприятия по предотвращению ЧС и порядок действия в случае возникновения ЧС

К пожарно-профилактическим мероприятиям можно отнести: правильный выбор оборудования и способы его монтажа, систематический контроль исправности защитных устройств на электрооборудовании, создание условий, обеспечивающих пожарную безопасность [23].

При обнаружении пожара необходимо выполнить следующий порядок действий:

1. Сообщить о пожаре в пожарную охрану по телефонам «112», «01» (с сотового тел. 101, 112).
2. Если нет опасности поражения электротоком, приступить к тушению пожара водой, или используйте плотную (мокрую ткань).
3. При опасности поражения электротоком отключить электроэнергию.
4. Горючие жидкости тушить водой нельзя (тушить песком, землёй, огнетушителем, если их нет, накрыть плотной смоченной в воде тканью).
5. При пожаре ни в коем случае не открывать форточки и окна.
6. Если не удаётся своими силами ликвидировать пожар, эвакуироваться и сообщить о пожаре работникам/соседям.
7. Встретить пожарных и провести их к месту пожара.

8. При высокой температуре, сильной задымлённости необходимо передвигаться ползком, так как температура у пола значительно ниже и больше кислорода.

Ликвидация пожара требует больших затрат. После пожара необходимо очистить помещение от мусора, устранить следы копоти с поверхностей, провести химчистку мебели, находящейся в помещениях, которые были задеты пожаром.

5.6 Вывод

В данном разделе дипломной работы были изложены требования к режиму труда и отдыха и рабочему месту сотрудника. Таким образом, сотрудник должен проводить не более 6 часов за компьютером, делая при этом перерывы на 15 минут каждый час.

Основными вредными факторами являются перенапряжение зрительного анализатора при длительной работе за ПК, превышение уровня шума, влияние электромагнитных излучений, отклонение от показаний микроклимата, недостаточная освещенность рабочей зоны. Соблюдение условий, позволяющих минимизировать опасные и вредные факторы, приведенные выше, позволит сохранить хорошую работоспособность в течение всего рабочего дня, повысит как в количественном, так и в качественном отношении производительность труда.

При анализе экологической безопасности было выявлено, что разработка оказывает влияние на литосферу. Для минимизации влияния необходимо следовать правилам утилизации отходов отработавшего оборудования.

Наиболее вероятными чрезвычайными ситуациями являются пожары, которые могут быть вызваны какой-либо неисправностью ПК. При возникновении пожара необходимо руководствоваться рекомендациями по предотвращению ЧС.

Заключение

В ходе работы над выпускной квалификационной работой была изучена литература и проанализированы полученные сведения, были выбраны нейросетевые модели для реализации – сверточные и рекуррентные LSTM нейронные сети, была спроектирована архитектура и реализовано программное обеспечение, была спроектирована структура выбранных нейросетевых моделей, обучены выбранные нейросетевые модели и проведены численные эксперименты на выборках при разных параметрах. В результате экспериментов были подобраны внутренние параметры моделей, показавших наилучшую точность классификации. Сверточная нейронная сеть по определению авторства достигла максимальной точности в 78%, а рекуррентная LSTM нейронная сеть 90%. Таким образом, LSTM сеть оказалась успешнее в решении задачи определения авторства по фрагменту произведения. Предварительная классификация текстов исходной выборки увеличила точность рекуррентной LSTM сети до 91%.

В результате работы было разработано программное обеспечение для определения автора по фрагменту произведения на основе искусственных нейронных сетей, способное определять автора текста с точностью до 91%, а также производить классификацию текстов с точностью до 90%.

Список публикаций не по тематике работы

1. Исаев Е. А. Классификация текста на изображениях реальных сцен при помощи сверточных нейронных сетей / Е. А. Исаев, Е. В. Михайлова, Ю. А. Иванова // Молодежь и современные информационные технологии : сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск. — Томск: Изд-во ТПУ, 2018. — [С. 66-67].

2. Кондратьева А.А., Маслов К.А., Михайлова Е.В. Искусственные нейронные сети в задачах повышения резкости изображений // Научная сессия ТУСУР – 2017: материалы Международной научно-технической конференции студентов, аспирантов и молодых ученых, посвященной 55-летию ТУСУРа. В 8 частях, Томск, 10-12 Мая 2017. – Томск: В-Спектр, 2017 – Т. 5 – С. 68 – 70

3. Кондратьева А.А., Маслов К.А., Михайлова Е.В. Искусственные нейронные сети в задачах повышения резкости изображений // Молодежь и современные информационные технологии: сборник трудов XV Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 4-7 Декабря 2017. – Томск: ТПУ, 2018 – С. 319 – 320

Список использованных источников и литературы

1. Автороведческая экспертиза [Электронный ресурс] / Википедия. – Режим доступа: https://ru.wikipedia.org/wiki/Автороведческая_экспертиза, свободный (дата обращения: 07.01.2019).
2. Батура Т.В. Формальные методы определения авторства текстов // Вестник новосибирского государственного университета. Серия: Информационные технологии, 2012. Том 10, выпуск 4.
3. Дроздова И. И., Обухова А. Д. Определение авторства текста по частотным характеристикам [Текст] // Технические науки в России и за рубежом: материалы VII Междунар. науч. конф. (г. Москва, ноябрь 2017 г.). — М.: Буки-Веди, 2017. — С. 18-21. — URL <https://moluch.ru/conf/tech/archive/286/13237/> (дата обращения: 06.01.2019).
4. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вести. МГУ. Сер. 9. Филология. 2000. JV* 2. С. 115-126. URL <http://www.philol.msu.ru/~lex/khmelev/published/vestnik/a.pdf> (дата обращения: 06.01.2019).
5. Метод опорных векторов [Электронный ресурс] / Википедия. – Режим доступа: https://ru.wikipedia.org/wiki/Метод_опорных_векторов, свободный (дата обращения: 07.01.2019).
6. Рысьмятова А.А. Использование сверточных нейронных сетей для задачи классификации текстов – Москва, 2016. – 34 с.
7. Рекуррентная нейронная сеть/ Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Рекуррентная_нейронная_сеть, свободный (дата обращения: 17.03.2019).
8. Understanding LSTM Networks/ colah's blog. – Режим доступа: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>, свободный (дата обращения: 17.03.2019).
9. Рекуррентная нейронная сеть в 10 строчек кода оценила отзывы зрителей нового эпизода “Звездных войн” [Электронный ресурс] /

Хабрахабр. – Режим доступа: <https://m.habr.com/ru/company/dca/blog/274027/>, свободный (дата обращения: 17.03.2019).

10. Gers F.A., Schmidhuber J. Recurrent Nets that Time and Count // Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, Volume 3, Pages: 189 – 194. URL <https://ieeexplore.ieee.org/document/861302> (дата обращения: 06.01.2019).

11. Word2Vec [Электронный ресурс] / Википедия. – Режим доступа: <http://ru.wikipedia.org/wiki/Word2Vec>, свободный (дата обращения: 07.01.2019).

12. Немного про word2Vec: полезная теория [Электронный ресурс] / NLPx – Tales of Data Science. – Режим доступа: <http://nlpx.net/language/ru/>, свободный (дата обращения: 07.01.2019).

13. TensorFlow API Documentation [Электронный ресурс]. – Режим доступа: https://www.tensorflow.org/api_docs/, свободный (дата обращения: 07.01.2019).

14. Исаев Е. А. Классификация текста на изображениях реальных сцен при помощи сверточных нейронных сетей / Е. А. Исаев, Е. В. Михайлова, Ю. А. Иванова // Молодежь и современные информационные технологии : сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск. — Томск: Изд-во ТПУ, 2018. — [С. 66-67].

15. Коэффициент скорости обучения (Learning rate) [Электронный ресурс]/ Loginom – Режим доступа: <https://wiki.loginom.ru/articles/learning-rate.html> свободный (дата обращения: 07.01.2019).

16. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы: Санитарно-эпидемиологические правила и нормы. – М.: Федеральный центр госсанэпиднадзора Минздрава России, 2003. – 54 с.

17. СН 2.2.4/2.1.8.562-96 Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки. Санитарные нормы: – М.: Информационно-издательский центр Минздрава России, 1997 г.

18. ГОСТ 12.1.002-84. Электрические поля промышленной частоты. – Введ. 01.01.86 – М.: Изд-во стандартов, 1984. – 7 с.

19. ГОСТ Р 55710-2013 Освещение рабочих мест внутри зданий. Нормы и методы измерений – Официальное издание. М.: Стандартинформ, 2016 год. – 20 с.

20. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений: Санитарные правила и нормы: – М.: Информационно-издательский центр Минздрава России, 2001. – 20 с.

21. ГОСТ Р 50948-2001 Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности: – М.: ИПК Издательство стандарт, 2002 год. – 11 с.

22. ГОСТ Р 55102-2012 Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов: – М.: Стандартинформ, 2014 год. – 16 с.

23. Действия при пожаре в квартире или здании [Электронный ресурс]/ SFU. – Режим доступа: <http://my.sfu-kras.ru/safety/fire-building/>, свободный (дата обращения: 07.04.2019).