TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа ядерных технологий
Направление подготовки 01.03.02 Прикладная математика и информатика
Отделение школы (НОЦ) Отделение экспериментальной физики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Тема работы
Сентимент-анализ в машинном обучении с использованием русскоязычных данных

УДК 004.85:004.934.2:811.161.1:59.942

Обучающийся

Группа	ФИО	Подпись	Дата
0В92	Кудинкина Екатерина Андреевна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ ИЯТШ	Семенов Михаил Евгеньевич	Кандидат ф. – м. наук, доцент		

Со-руководитель ВКР (по разделу «Концепция стартап-проекта»)

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент Бизнес школа ТПУ	Таран Екатерина Александровна	Кандидат э.наук, доцент		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Антоневич Ольга Алексеевна	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП/ОПОП, должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ ИЯТШ	Крицкий Олег Леонидович	Кандидат ф. – м. наук, доцент		

Томск – 2023 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа ядерных технологий
Направление подготовки 01.03.02 Прикладная математика и информатика
Отделение школы (НОЦ) Отделение экспериментальной физики

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Крицкий О.Л.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы (стартап)

Студенту:

Группа	ФИО
0В92	Кудинкиной Екатерине Андреевне

Тема работы:

Сентимент-анализ в машинном обучении с использованием русскоязычных данных	
Утверждена приказом директора (дата, номер)	

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	*Набор данных из заголовков новостей, связанных с с акционерными компаниями*
Перечень подлежащих исследованию, проектированию и разработке вопросов	*1. Провести сравнительный анализ методов сопоставления критерия эмоций с входными данными* *2. Провести сравнительный анализ методов машинного обучения* *3. Выбор методов предобработки данных и обучения классификатора* *4. Разработать алгоритм и программную реализацию для машинного обучения* *5. Провести тестирование и провести оценку качества полученного классификации*
Перечень графического материала	*1. Графики метрики точности относительно количества эпох при обучении нейронной сети* *2. График распределения классификации выборки*
Консультанты по разделам выпускной квалификационной работы *(если необходимо, с указанием разделов)*	
Раздел	**Консультант**
Концепция стартап-проекта	Таран Екатерина Александровна, доцент ШИП
Социальная ответственность	Антоневич Ольга Алексеевна, доцент ООД ШБИП

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ОЭФ	Семенов Михаил Евгеньевич	к. ф.-м. н., доцент		

Задание принял к исполнению студент:

Группа	**ФИО**	**Подпись**	**Дата**
0В92	Кудинкина Екатерина Андреевна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО		
0В92	Кудинкина Екатерина Андреевна		
Школа	**Инженерная школа ядерных технологий**	**Отделение (НОЦ)**	**Экспериментальной физики**
Уровень образования	Бакалавриат	**Направление/ специальность**	01.03.02 Прикладная математика и информатика

Тема ВКР:

Сентимент-анализ в машинном обучении с использованием русскоязычных данных	
Исходные данные к разделу «Социальная ответственность»:	
Введение − Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. − Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* программное обеспечение *Область применения:* мониторинг и контроль информации об опционах *Рабочая зона:* офисное помещение *Размеры помещения:* 27,2 м^2 *Количество и наименование оборудования рабочей зоны:* 2 персональных компьютера *Рабочие процессы, связанные с объектом исследования, осуществляемые в рабочей зоне:* алгоритмическая и программная разработка с использованием персонального компьютера
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения: - специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; - организационные мероприятия при компоновке рабочей зоны.	– ГОСТ 12.2.032-78 "Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования"; –Трудовой кодекс Российской Федерации: федер. Закон от 30 дек. 2001 г. №197-ФЗ Раздел 10; –РД 153-34.0-03.298-2001 «Типовая конструкция по охране труда для пользователей персональными электронно-вычислительными машинами (ПЭВМ) в электроэнергетике»
2. Производственная безопасность при разработке проектного решения: - Анализ выявленных вредных и опасных производственных факторов	**Опасные факторы:** – Опасность поражения электрическим током; **Вредные факторы:** – Отклонение показателей микроклимата; – Недостаточная освещенность рабочей зоны; – Пониженная световая и цветовая контрастность; – Повышенный уровень шума на рабочем месте; – Повышенный уровень статического электричества; **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** изоляция электрических проводов, устройства защитного заземления и зануления.

3. Экологическая безопасность при разработке проектного решения:	**Воздействие на селитебную зону:** утилизация компьютеров, их составляющих, компьютерных аксессуаров; **Воздействие на литосферу:** образование отходов в ходе поломки и утилизации компьютеров, оргтехники и бумаги;
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения:	**Возможные ЧС:** пожар; **Наиболее типичная ЧС:** пожар;
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Антоневич Ольга Алексеевна	к.б.н., доцент		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В92	Кудинкина Екатерина Андреевна		

ЗАДАНИЕ К РАЗДЕЛУ
«КОНЦЕПЦИЯ СТАРТАП-ПРОЕКТА»

Обучающемуся:

Группа	ФИО
0В92	Кудинкина Екатерина Андреевна

Школа	ИЯТШ	Отделение школы (НОЦ)	ОЭФ
Уровень образования	Бакалавриат	Направление/ООП/ОПОП	01.03.02 Прикладная математика и информатика

Перечень вопросов, подлежащих разработке:	
Проблема конечного потребителя, которую решает продукт, который создается в результате выполнения НИОКР (функциональное назначение, основные потребительские качества)	*Выявить проблему, которую будет решать данное программное обеспечение*
Способы защиты интеллектуальной собственности	*Рассмотреть способы защиты интеллектуальной собственности для нашего продукта*
Объем и емкость рынка	*Рассчитать примерный объем рынка с помощью методов "сверху-вниз" и "снизу-вверх"*
Современное состояние и перспективы отрасли, к которой принадлежит представленный в ВКР продукт	*Оценить возможности данного продукта относительно его отрасли, найти перспективы развития*
Себестоимость продукта	*Посчитать примерную себестоимость продукта*
Конкурентные преимущества создаваемого продукта и Сравнение технико-экономических характеристик продукта с отечественными и мировыми аналогами	*Исследовать подобные решения на рынке, рассмотреть преимущества и недостатки*
Целевые сегменты потребителей создаваемого продукта	*Рассмотреть потребительскую аудиторию разрабатываемого решения*
Бизнес-модель проекта, производственный план и план продаж	*Построить модель Остервальдера*
Стратегия продвижения продукта на рынок	*Продумать способы продвижения данного продукта на рынок*
Перечень графического материала:	
При необходимости представить эскизные графические материалы (например, бизнес-модель)	*Бизнес-модель - 1 шт., таблица - 7 шт.*

Дата выдачи задания к разделу в соответствии с	

календарным учебным графиком	

Задание выдал консультант по разделу «Концепция стартап-проекта» (со-руководитель ВКР):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ШИП	Таран Екатерина Александровна	к. э. н., доцент		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
0В92	Кудинкина Екатерина Андреевна		

Реферат

Выпускная квалификационная работа содержит 65 страниц, 17 таблиц, 18 рисунков, 22 источника.

Ключевые слова: машинное обучение, сентимент-анализ, глубокое обучение, языковые модели, обработка естественного языка.

Цель работы- провести сентимент-анализ в машинном обучении с использованием русскоязычных данных.

Данная квалификационная работа посвящена разработке программного обеспечения для анализа тональности русскоязычных новостей компаний, которые имеют опционы на бирже. Основная часть работы посвящена модулю реализации различных языковых моделей, используемых в анализе тональности.

Для обработки были взяты данные Financial News Sentiment Dataset (FiNeS), содержащих в себе заголовки финансовых новостей о компаниях, торгующихся на Московской и Санкт-Петербургской биржах. Сбор новостей проводился из RSS-лент следующих источников: РБК, Коммерсантъ, Финам, Investing, и Ведомости. Для каждой новости автоматически проверялось наличие наименования определённой компании в заголовке. Новости, без упоминания компаний не были включены в набор данных.

Для разработки программного модуля был использован язык программирования Python, работа проводилась в Jupiter Notebook и во встроенном редакторе Google Colaboratory.

Оглавление

Определения, сокращения, обозначения, нормативные ссылки

Датасет - обработанный и структурированный массив данных.

NLTK - пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python.

NLP - это технология машинного обучения, которая дает компьютерам возможность интерпретировать, манипулировать и понимать человеческий язык.

EDA - разведочный анализ данных.

Hugging Face Transformers - конвейер с языковыми моделями, которые можно использовать для обучения.

Tensorflow - открытая программная библиотека для машинного обучения, разработанная компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов, достигая качества человеческого восприятия.

Стоп-слово - слова, которые присутствуют в любом языке и не привносят никакой полезной информации в предложение.

Word2Vec - общее название для совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке.

BOW - это способ представления текстовых данных при моделировании в машинном обучении.

Введение

Сентимент-анализ - это мощное приложение машинного обучения, которое включает в себя анализ и категоризацию мнений, эмоций и отношения людей к различным темам, продуктам или услугам, выраженных в письменной или устной форме. С быстрым ростом социальных сетей, онлайн-обзоров и отзывов клиентов анализ настроений стал важным инструментом для бизнеса, позволяющим извлекать ценную информацию из огромных объемов неструктурированных данных. Он включает в себя использование методов обработки естественного языка (natural language processing, NLP) и алгоритмов машинного обучения для автоматического определения и классификации эмоций как положительных, отрицательных или нейтральных [1]. Анализ эмоций имеет множество применений, начиная от исследования рынка, управления брендом, обслуживания клиентов и политического анализа.

В данной работе мы рассмотрим основы обработки настроений с использованием машинного обучения, приложения и классические методы, используемые в предметной области.

В литературе выделяют два основных подхода к анализу настроений в машинном обучении: на основе правил и на основе машинного обучения.

Подходы, основанные на правилах, включают использование набора предопределенных правил или словарей для определения полярности тональности текста [2]. Эти правила или лексиконы состоят из списков слов и фраз, которые ассоциируются с положительными или отрицательными эмоциями. Тональность фрагмента текста определяется путем подсчета количества положительных и отрицательных слов в тексте и сравнения подсчетов.

С другой стороны, подходы, основанные на машинном обучении, включают обучение модели машинного обучения на размеченном наборе данных текста с известной полярностью настроений [3]. Модель обучается определять закономерности и взаимосвязи между словами и тональностью и использует эти знания для прогнозирования тональности нового текста. Популярные алгоритмы

машинного обучения, используемые для анализа настроений, включают логистическую регрессию, наивный Байесовский классификатор,

Анализ настроений имеет множество применений в различных областях, включая исследования биржи. Благодаря ему, появляется возможность объективно оценить то или иное событие, связанное с желаемым опционом и ускорить принятие решения. Однако, интеллектуальная система запросто может столкнуться с переобучением и это уже негативно влияет на изучение опционной биржи, в которой важна скорость и взвешенность принятия решения.

Целью данной работы является исследование различных языковых моделей, направленных на сентимент-анализ финансовых новостей, и выбор наиболее точно обученной модели на основе критериев машинного обучения. Для достижения цели необходимо выполнить следующие задачи:

- Реализовать алгоритм предварительной обработки данных
- Выполнить анализ данных, используя стандартные модели классификации типа “Bag of words”, “word2vec”.
- Построить несколько разных методов и техник в текстовом классификаторе: от самого простого до глубокого обучения, рассмотреть основные метрики оценки обучения модели.
- Выбрать наиболее точно обученную модель, опираясь на основные критерии оценивания качества обучения.

1. Теоретическая часть

1.1. Препроцессинг текста

Предварительная обработка данных в машинном обучении – это важный шаг, который помогает повысить качество данных. Предобработка данных в машинном обучении относится к технике подготовки необработанных данных с целью сделать их пригодными для построения и обучения моделей машинного обучения. Иными словами, это метод интеллектуального анализа данных, который преобразует необработанные данные в понятный и читаемый формат [9].

Предварительная обработка данных является одним из основных этапов, от качества выполнения которого зависит получение качественных результатов процесса анализа данных. Без подготовки данных не обходится ни один нейросетевой метод. Как правило, при описании различных нейроархитектур предполагается, что данные для обучения уже представлены в том виде, в котором требует нейросеть, однако на практике дела обстоят совсем иначе, именно этап предобработки данных может занимать большую часть времени, отведенного на проект в целом [10]. Результат обучения нейросети также может зависеть от того, в каком виде представлена информация для ее обучения. Таким образом, предварительная обработка данных позволяет повысить качество как интеллектуального анализа данных, так и самих данных. Предобработка данных состоит из этапов [1]:

− очистка данных, которая направлена на повышение качества данных за счет присваивания пропущенных значений и удаления выбросов;

− сокращение объема данных, которое уменьшает объем данных и, следовательно, снижает связанные с ними вычислительные мощности;

− масштабирование данных – направлено на преобразование исходных данных в аналогичные диапазоны для прогнозного моделирования;

− преобразование, целью которого является организация исходных данных в подходящие форматы для различных алгоритмов интеллектуального анализа данных;

1.2 Мешок слов (Bag of words)

Мешок слов - способ классифицировать данные, опираясь на характер появления слов в датасете. Она учитывает построение словаря слов из исходных данных, в котором генерируется список уникальных слов и показывается количество появления этих слов в таблице [5]. Модель имеет такое название потому что в ней не имеет значение порядок расположения слов в предложении, удаляется вся структура строки. Остаются только слова, которые появляются в любом порядке, а сама модель проверяет появление слов из созданного уникального словаря.

Метод тесно связан с TF-IDF (term frequency-inverse document frequency) моделью. Рассмотрим подробнее формулу, по которой описывается этот алгоритм.

TF- частота слов, вычисляется как отношение количества появления слова в датасете ко всей сумме слов:

$$TF(t,d) = \frac{n_t}{\sum_k n_k}, \qquad (2)$$

где n_t - количество появлений слова в датасете, а n_k - общее количество слов.

IDF - обратная частота слов, показывает коэффициент важности определенного слова. У каждого уникального слова один коэффициент и свой. Он рассчитывается как:

$$IDF(t,D) = log\frac{|D|}{\left|\left\{d_i \in D |\ t \in d_i\right\}\right|}, \quad (3)$$

где D - количество всех строк, а $\left\{d_i \in D |\ t \in d_i\right\}$ - количество строк, в которых встречается слово.

В конечном итоге мы перемножаем эти два показателя и получаем итоговый вес. Данный метод помогает отсеять предлоги, местоимения, т.е. слова, который не особо влияют на смысловую нагрузку.

1.3 Word2vec

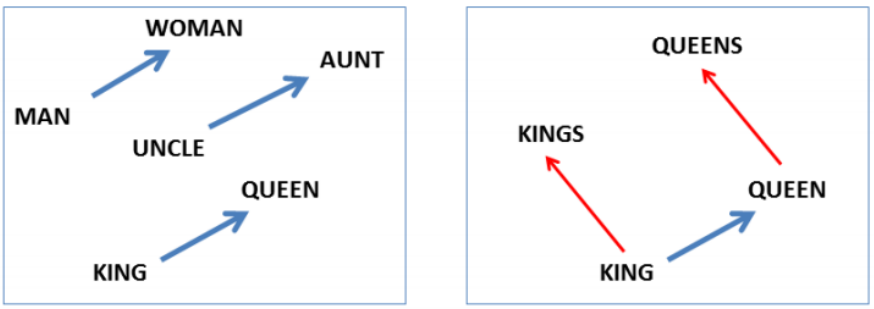
Кодировка данных может производиться разными способами, самый популярный- создать словарь из всех слов датасета и присвоить им порядковый номер. Главный минус - отсутствие смысловой нагрузки, нужно показать семантическую близость [5]. В 2013 году появилась модель word2vec, которая находит вектор слова по соседним словам. Математическое описание модели выглядит следующим образом:

$$P(w_0|w_1) = \frac{e^{s(w_0,w_1)}}{\sum w_1 \in V e^{s(w_t,w_1)}} \quad , (1)$$

где w_0 - ключевое слово, представленное в виде вектора, w_1 - вектор слова, окружающего целевое слово, которое рассчитано при помощи усреднения. $s(w_0|w_1)$- функция, которая ставит число в соответствие паре векторов. Эта функция может иметь вид меры сходства между двумя ненулевыми векторами внутреннего пространства произведений, которое измеряет косинус угла между ними [5]. Данную формулу можно представить в дифференцированном виде и присвоить метод обратного распространения ошибки. В основе процесса создания алгоритма лежит следующий подход: есть слово, вектор которого мы хотим вычислить, для этого берем (2k+1) слов, окружающих целевое, последовательно друг за другом. Они будут отвечать за контекст, длиной k в каждую сторону от центра. Каждому из этих слов будет сопоставлен уникальный вектор, полученный путем усреднения [5].

Данный способ аналогичен методу BOW, единственное различие в том, что алгоритм получает наборы слов последовательно, то есть важным является порядок, с которым обучается инструмент, но расположение самих слов внутри последовательности роли не играет [6]. Важное достоинство модели word2vec - она способна вычислять семантику, чего не могут большинство моделей. Эта способность не включена в саму модель, но если

хорошо ее обучить, то она покажет контекст употребляемых слов. Краткая характеристика метода хорошо продемонстрирована на схеме внизу [6]:



(Mikolov et al., NAACL HLT, 2013)

Рисунок 1 - Схема связей между словами по методу word2vec

Способ преобразования данных в числовые вектора называется эмбеддингом.

1.4 Модель логистической регрессии

Так как мы имеем дело с тремя классами, на которые нужно соотнести входные данные, то используем мультиномиальную логистическую регрессию. Она является самой быстрой для обучения моделью, особенно с разреженными данными (в нашем случае в датасете больше половины строк относятся к нейтральным). В биномиальной логистической регрессии мы используем сигмоидную функцию в качестве обучения, здесь берется функция вида softmax для K классов:$softmax(x_i = \text{c}) = \frac{e^{x_i}}{\sum_{K=1}^{k} e^{z}j}$, а вероятность будет равна: $Pr(Y = c|\overline{X} = x) = \frac{e^{wx+b}}{\sum_{K=1}^{k} e^{wx+b}}$. В данных формула c - константа, выполняет роль вектора коэффициента регрессии, который устанавливает разницу между другими векторами и вектором, рассматриваемым по порядку.

1.5 Наивный Байесовский классификатор

Данная модель построена с опорой на формулу Байеса, которая имеет следующий вид: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$(4), где $P(A|B)$ - вероятность

наступления события B при происходящем событии A. $P(B|A)$ показывает частоту происхождения события B при наступлении события A. Оставшиеся вероятности $P(A)$ и $P(B)$ демонстрируют вероятности независимости от других событий. Суть алгоритма в том, что мы относим данные к определенным классам по векторам признаков, но допуская, что признаки не коррелируют друг с другом. Сам классификатор не нуждается в большой выборке и при этом до сих пор остается с хорошими метриками качества обучения.

В основе вероятностной модели лежит множество событий (заголовков) $x = (x_1,..., x_i)$. Метод соотносит к каждому событию условную вероятность, равную $P(C_k|x_i)$, C_k - класс, на который классифицируем. Подставляем в формулу Байеса (4), предполагаем что признаки независимы, и получаем $P(C_k|x_i) = \prod_i P(x_i|C_k)$ - байесовский классификатор [7]. Он присваивает каждому событию факт принадлежности к определенному классу, имеющий вид $y = arg_k max_{1..k} \prod_n P(x_i|C_k)$. Класс y выбирается таким образом, чтобы максимизировать функцию правдоподобия, которая представляет собой произведение условных вероятностей признака x_i. Наивный Байес предсказывает класс с наибольшей условной вероятностью для вектора признаков x.

1.6 LSTM с использованием word2vec

Перейдем к использованию нейронных сетей, одна из таких - LSTM (Long Short Term Memory), появившаяся в конце 90-х. Она относится к разновидности RNN (рекуррентная нейронная сеть), где связи между нейронами создают направленную последовательность [8]. Изначально RNN стала популярна благодаря своей способности запоминать с помощью внутренней памяти. Это позволяет выявить тренд, найти закономерность, хорошо коррелирует с нашей выборкой потому что при неоднократном

упоминании компании, опционы которой торгуются на бирже, можно найти связь. LSTM стала некоторым улучшением RNN, так как она научилась учитывать предыдущие события при обучении. Это происходит с помощью блока памяти или “индикатора ячейки”, которая принимает решение что делать с полученной информацией: удалить или сохранить. На это решение влияют функции с весами, которые меняются при обучении путем обратного распространения ошибки [8].

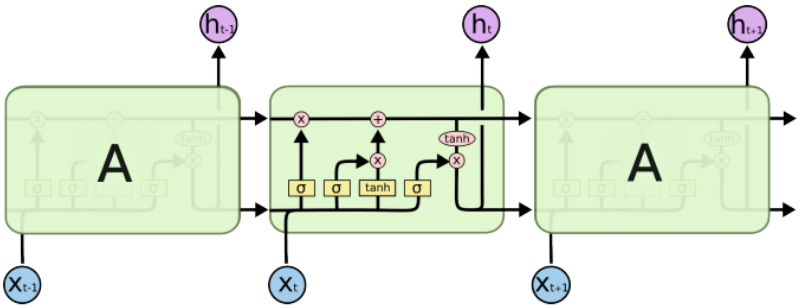


Рисунок 2 Повторяющий модуль LSTM сети, внутри которого 4 взаимодействующих слоя [9]

1.7 BERT

BERT – это двунаправленная мультиязычная модель с transformer-архитектурой (рис. 1), предназначенная для решения конкретных NLP-задач, например, определение эмоциональной окраски (тональности) текста, вопросно-ответные системы, классификация текстов, построение выводов по тексту и т.д. Помимо распознавания речи, классической NLP-задачей является анализ текста, включая извлечение данных, информационный поиск и анализ высказываний. Также к обработке естественного языка относятся генерирование текстов, синтез речи, машинный перевод и автоматическое реферирование, аннотирование и упрощение текстовой информации. Таким образом, цель применения NLP-технологий – это не только распознавание живого языка средствами искусственного интеллекта, но и возможность адекватного с ним взаимодействия. Последнее, фактически, означает понимание AI-инструментом устной или письменной речи [2].

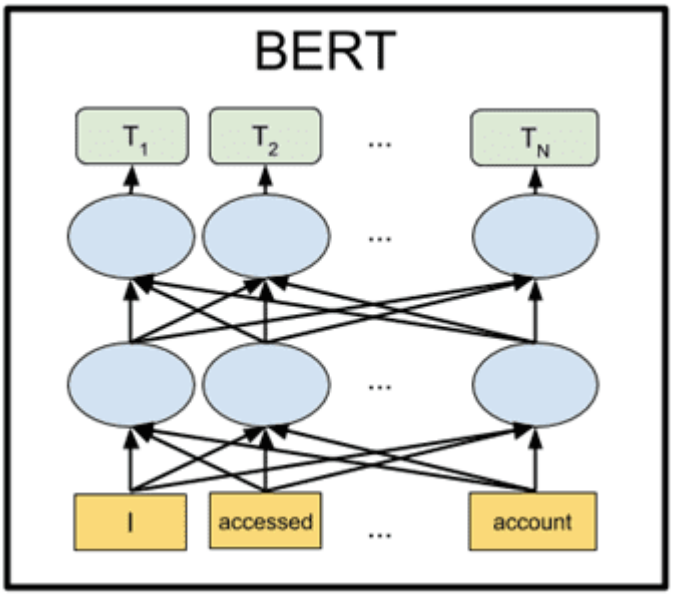


Рисунок 3 - Архитектура BERT

BERT работает по принципу векторного представления слов, основанном на контекстной близости, когда слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты векторов. Полученные векторы могут быть использованы для обработки естественного языка и машинного обучения, в частности, для прогнозирования слов [8].

Чтобы натренировать BERT на предсказывание слов, на вход нейросети подаются фразы, где часть слов заменена на маску [MASK]. При подаче текста на вход BERT-модели сначала выполняется его токенизация — разбиение на более мелкие единицы (токены): абзацы делятся на предложения, предложения на слова и пр. Входной текст разбивается на список токенов, доступных в словаре [8].

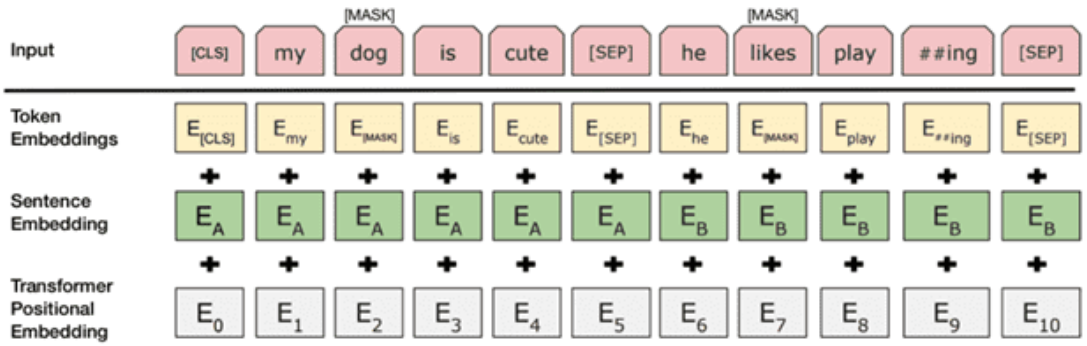


Рисунок 4 - Токенизация в BERT [8]

1.8 ROBERT

Представленный на Facebook надежно оптимизированный подход BERT RoBERTa, представляет собой переподготовку BERT с улучшенной методикой обучения, на 1000% больше данных и вычислительной мощностью.

Чтобы улучшить процедуру обучения, RoBERTa удаляет задачу «Предсказание следующего предложения» (NSP) из предварительной подготовки BERT и вводит динамическое маскирование, чтобы замаскированный токен изменялся в эпоху обучения. Также было обнаружено, что большие размеры обучения партии более полезны в процедуре обучения.

Важно отметить, что RoBERTa использует 160 ГБ текста для предварительного обучения, в том числе 16 ГБ Books Corpus и английскую Википедию, используемые в BERT. Дополнительные данные включены CommonCrawl News набор данных (63 миллиона статей, 76 ГБ), корпус веб-текста (38 ГБ) и «Истории из общего сканирования» (31 ГБ). Это вкупе с колоссальным временем работы 1024 графических процессоров V100 Tesla в течение дня привело к предварительной подготовке RoBERTa.

2. Практическая часть

2.1 Предварительная обработка данных

2.1.1 Разведочный анализ данных

Разведочный анализ данных (EDA) - анализ основных свойств данных, позволяющий найти общие зависимости или уточнить их свойства. В данной работе используется готовый датасет, который представлен в виде таблицы с колонками “Заголовок”, “Ссылка на новость”, “Краткое описание новости”, “Дата публикации”, “Билет”, “Коэффициент тональности” [4]. Для анализа тональности нам понадобится всего 2 колонки: “Заголовок” и “Коэффициент тональности”, остальные данные не пригодятся. Для этого мы удалили другие колонки, а коэффициент тональности отнесли в три группы: “позитивный”, “нейтральный” и “негативный”. Условие соотнесения в группу “нейтральный” было следующим: коэффициент тональности был меньше 0.05 и больше -0.05. Мы получили таблицу, состоящую из 2 колонок и 532 строк. Построим график частоты появления заголовков по тональности:

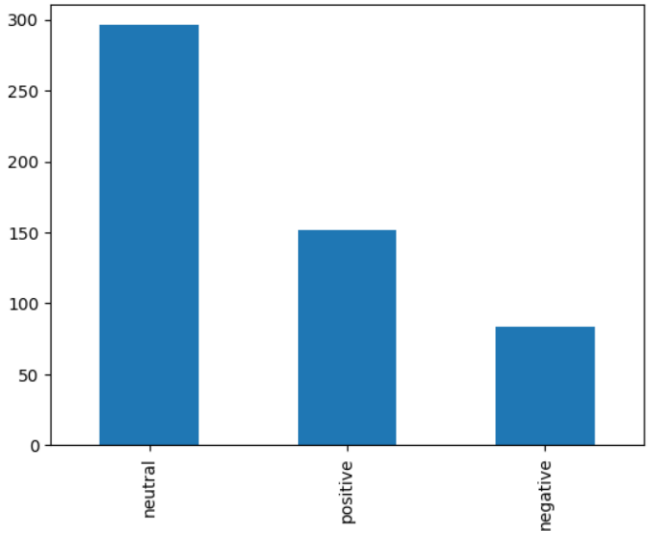


Рисунок 5 - График частоты появления заголовка, определенного по тональности

Судя по графику, мы получили преобладающее количество нейтральных новостей, которые не будут влиять на опцион прозвучавшей в заголовке компании. Для последующей работы мы отдельно сохраним наш

датасет в бинарной системе счисления: позитивный представлен в виде 1, нейтральный в виде 0, негативный в виде -1.

2.2 Предобработка датасета

Зачастую, на входе мы получаем данные, в которых много “шума”: верхние регистры, лишние пробелы, знаки препинания, пустые слова. Нужно предварительно очистить датасет и подготовить его для дальнейшего использования. Это повысит качество обучения, уменьшит ненужную информацию [5]. В нашей работе мы выполнили токенизацию (разделение слов на отдельные единицы представления - токены). Этот процесс был проведен при помощи библиотеки NLTK, встроенной в язык программирования Python. С помощью этой же библиотеки мы удалили стоп-слова (пустые токены, не несущие смысловой нагрузки и мешающие обучению классификатора). Далее удалили верхние регистры, пустые пробелы, знаки препинания. Импортировали инструмент pymorphy2, позволяющий проводить лемматизацию - приведение слова в начальную форму. Все эти действия описываются термином “препроцессинг”, его необходимо в любом анализе данных, так как это сильно влияет на результат обучения.

Preprocessed_texts	Sequences
[электромобильный, стартап, arrival, экс-глава...	[610, 358, 611, 234, 612, 88, 5]
[экс-глава, нмтп, рассказать, напрячь, отношен...	[234, 613, 359, 614, 360, 615, 616]
[шрёдёр, отклонить, предложение, войти, совет,...	[617, 173, 174, 618, 89, 76, 2]

Рисунок 6 - Пример предобработки данных для достижения наилучшего результата обучения классификатора

2.3. Классификация набора данных

Для того, чтобы извлечь полезную информацию из датасета, необходимо его преобразовать в полезную функцию. Компьютер не может работать с неструктурированными данными, поэтому необходимо представить датасет в виде векторов чисел. Существует большое количество способов представления информации, в нашей работе мы будем использовать BOW, word2vec.

2.3.1 Word2vec

На языке Python данная модель имеет следующий вид: word2vec(size=300, min_count = 1). Здесь size - это размерность векторов, полученных из слов выборки, а min_count регулирует слова с наименьшей частотой появления, т.е. пропускает их. В самой модели существует еще множество параметров, подлежащих регулированию, но у нас выборка сравнительно небольшая и не нуждается в них, иначе возрастает риск пропустить важную информацию. Далее просто обучаем модель, используя метод train и сохраняем полученные вектора в экземпляре KeyedVectors в переменную word2vec_model.

```
{'поощрялась': array([-3.5042786e-03,  5.2169146e-04,  2.1134880e-03,  3.4255565e-03,
         2.9535536e-03, -2.5336174e-04,  3.5182014e-03,  4.7578118e-03,
        ...........................................................
        -2.3358944e-03,  3.9440244e-03,  7.6886819e-05, -3.4618229e-04],
       dtype=float32),
```

Рисунок 7 - Пример представления слова из словаря в виде вектора, с помощью модели word2vec.

2.3.2 Мешок слов (Bag of words)

В нашем случае мы рассмотрели частоту появления определенных слов, представили их в виде векторов. Каждое из слов, показанных на рис. 1.2, повторялось 1 раз.

```
['электромобильный',
 'стартап',
 'arrival',
 'экс-глава',
 'yota',
 'уйти',
 'россия']
```

Рисунок 8 - Словарь уникальных слов, составленный для модели классификации “BOW”

Далее мы применяем векторизацию ко всему датасету, получаем представление токенов в виде векторов. Количество векторов в нашем случае достигает 1538. Данное представление можно использовать для обучения классификатора.

2.4. Модели для определения тональности заголовков новостей в машинном обучении

После того, как мы провели препроцессинг данных, представили их в качестве векторов по нескольким методам, таким как: BOW и word2vec, мы можем проводить обучение классификатора. Рассмотрим различные модели - от тривиальной до модели, обученной с помощью deep learning.

Для единого представления всех моделей, мы создали функцию классической модели, в которой разделили выборку на тренировочную и тестовую, добавили основные метрики по оценки качества обучения модели. Метрики такие, как: precision, accuracy, f-score, recall, macro-averaged, weighted-averaged. Пройдемся кратко по их определениям:

- precision показывает верность правильно соотнесенных ответов к классу
- recall показывает насколько верно модель отнесла на классы данные
- accuracy рассчитывает точность общего предсказания на классы относительно правильного ответа
- f-score средняя гармоническая мера, объединяющая precision и recall
- macro-averaged рассчитывает среднее относительно количества классов
- weighted-averaged показывает среднее относительно классов, пропорционально количеству к принадлежности каждого из них

2.4.1 Модель логистической регрессии

Мы провели обучение, используя “BOW”. Ниже представлены полученные метрики:

	precision	recall	f1-score
-1	0.50	0.06	0.11
0	0.55	0.83	0.66
1	0.23	0.10	0.14
accuracy			0.51
macro avg	0.43	0.33	0.30
weighted avg	0.45	0.51	0.43

Рисунок 9 - Результаты обучения, используя мультиномиальную логистическую регрессию

Метрики обученной модели не очень хорошие, хуже всего машина распознает положительные заголовки, лучше всего, как и ожидалось, нейтральные. Неудовлетворительна метрика полноты (recall), это значит что машина неверно находит долю заголовков, имеющих негативный характер. Общая точность распознавания равна 51%.

2.4.2 Наивный Байесовский классификатор

Мы провели обучение на выборке, преобразованной в “BOW”, используя наивный Байесовский классификатор и получили следующие метрики:

	precision	recall	f1-score
-1	0.43	0.29	0.35
0	0.74	0.84	0.78
1	0.73	0.66	0.70
accuracy			0.70
macro avg	0.63	0.60	0.61
weighted avg	0.69	0.70	0.69

Рисунок 10 - Метрики качества классификатора методом мультиномиального наивного Байеса

Метрики определения негативного заголовка сильно страдают, делаем предположение что это происходит из-за недостаточности данных или вышеперечисленные модели плохо подходят для нашей задачи. Из главных плюсов можно отметить быструю обучаемость модели, скорость работы с данными, маленький объем памяти.

2.4.3 LSTM с использованием word2vec

В данном случае мы используем преобразованные вектора с помощью модели word2vec, уравниваем их по максимально длинному заголовку: заполняем пустые пробелы нулями. Далее создаем саму нейронную сеть, состоящую из 3 слоев:

```
lstm_m = Sequential()
lstm_m.add(Embedding(max_words, 300, input_length=maxlen))
lstm_m.add(LSTM(32, recurrent_dropout = 0.2))
lstm_.add(Dense(1, activation='sigmoid'))
```

Архитектура нейронной сети: создаем ячейку, в которой хранится сеть. Во второй строке мы создаем первый слой, принимающий на вход плотное векторное представление слов. В третьей строке появляется второй слой с LSTM, в которой 32 ячейки. Данный слой исключает некоторый процент нейронов для того, чтобы исключить возможное переобучение нейронной сети [9]. И в четвертой строке завершающий слой с одним полносвязным нейроном, функция активации - сигмоидная. Последний слой используется для классификации.

```
lstm_m.compile(optimizer='adam',
               loss='categorial_crossentropy',
               metrics=['accuracy'])
```

Компилируем модель со следующими параметрами: функция оптимизатора - adam, функция ошибки - категориальная перекрестная энтропия, основная метрика - точность.

Далее обучаем нейронную сеть на выборке, задаваемые параметры:

- epochs = 15
- batch_size = 128

Первый параметр определяет количество прохождений нейронной сети по датасету туда и обратно, трудно определить точное число прохождений для обучения. Если будет слишком много прохождений, то модель переобучится, а одного прохождения недостаточно для полного обучения. Второй параметр влияет на устойчивость сходимости датасета, нужно регулировать его относительно возможностей персонального компьютера.

```
Epoch 1/15
176/176 [==============================] - 171s 929ms/step - loss: 0.5473 - accuracy: 0.7182 - val_loss: 0.3482 - val_accuracy: 0.8592
Epoch 2/15
176/176 [==============================] - 151s 856ms/step - loss: 0.2852 - accuracy: 0.8873 - val_loss: 0.3079 - val_accuracy: 0.8764
Epoch 3/15
176/176 [==============================] - 145s 825ms/step - loss: 0.2108 - accuracy: 0.9238 - val_loss: 0.3274 - val_accuracy: 0.8820
Epoch 4/15
176/176 [==============================] - 139s 789ms/step - loss: 0.1728 - accuracy: 0.9380 - val_loss: 0.3348 - val_accuracy: 0.8788
Epoch 5/15
176/176 [==============================] - 137s 781ms/step - loss: 0.1520 - accuracy: 0.9482 - val_loss: 0.3513 - val_accuracy: 0.8796
Epoch 6/15
176/176 [==============================] - 139s 787ms/step - loss: 0.1321 - accuracy: 0.9550 - val_loss: 0.3680 - val_accuracy: 0.8732
Epoch 7/15
176/176 [==============================] - 136s 773ms/step - loss: 0.1076 - accuracy: 0.9647 - val_loss: 0.4069 - val_accuracy: 0.8596
Epoch 8/15
176/176 [==============================] - 137s 780ms/step - loss: 0.0983 - accuracy: 0.9677 - val_loss: 0.3998 - val_accuracy: 0.8596
Epoch 9/15
176/176 [==============================] - 137s 779ms/step - loss: 0.0843 - accuracy: 0.9734 - val_loss: 0.5172 - val_accuracy: 0.8576
Epoch 10/15
176/176 [==============================] - 136s 771ms/step - loss: 0.0764 - accuracy: 0.9759 - val_loss: 0.4945 - val_accuracy: 0.8672
Epoch 11/15
176/176 [==============================] - 135s 768ms/step - loss: 0.0625 - accuracy: 0.9819 - val_loss: 0.5236 - val_accuracy: 0.8608
Epoch 12/15
176/176 [==============================] - 137s 778ms/step - loss: 0.0622 - accuracy: 0.9815 - val_loss: 0.5791 - val_accuracy: 0.8688
Epoch 13/15
176/176 [==============================] - 135s 769ms/step - loss: 0.0586 - accuracy: 0.9820 - val_loss: 0.6478 - val_accuracy: 0.8652
Epoch 14/15
176/176 [==============================] - 135s 767ms/step - loss: 0.0538 - accuracy: 0.9836 - val_loss: 0.6379 - val_accuracy: 0.8564
Epoch 15/15
176/176 [==============================] - 136s 770ms/step - loss: 0.0467 - accuracy: 0.9865 - val_loss: 0.5873 - val_accuracy: 0.8592
```

Рисунок 11 - Эпохи прохождения языковой модели LSTM по датасету

В среднем одно прохождение по epoch занимало около 2 минут, изначально доля правильной классификации была 0,71 и уже к третьей эпохе она составляла 0,92. Качество обучения модели на проверочной выборке незначительно снизилось, значит машина начала переобучаться.

Посмотрим как графически меняется доля точности верной классификации относительно двух выборок: обучающей и тренировочной.

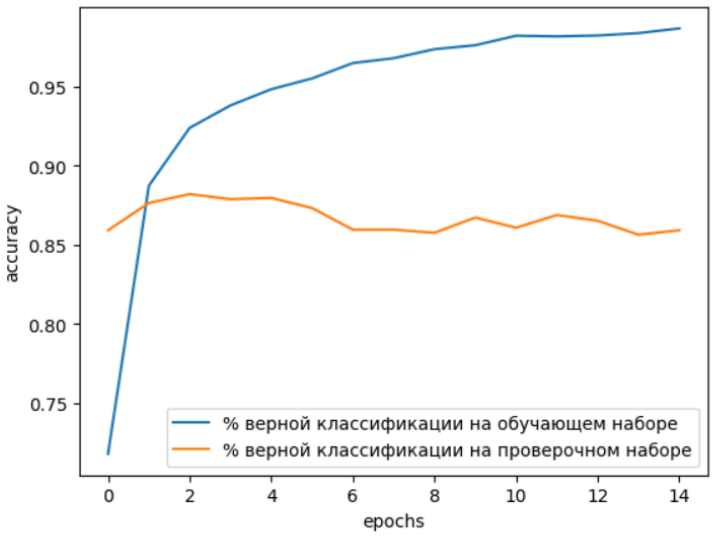


Рисунок 12 - График соотношения точности к эпохам на основе сравнения с обучающей выборкой и проверочной

На графике видно, что показатель точности на обучающей выборке растет, а на тренировочной падает. С одной стороны наша нейронная сеть с легкостью может классифицировать обучающую выборку с высокой точностью, а с другой стороны она переобучивается на тренировочной выборке. Посмотрим показатель точности работы языковой модели на тестовых данных:

```
accuracy = lstm_m.evaluate(x_test, y_test, verbose=1)
```

```
782/782 [==============================] - 49s 62ms/step - loss: 0.6529 - accuracy: 0.8424
```

Рисунок 13 - Проверка работы модели LSTM с тестовой выборкой

Точность классификации составила 0,8424, меньше чем на проверочной выборке.

2.4.4 BERT модель

Загружаем модель “BERTforSequenceClassification” вместе с встроенным токенизатором (так как мы хотим преобразовать заголовки с таким же словарем, с каким была предобучена данная языковая модель). Мы берем модель ‘bert-base-uncaused’, так как хотим чтобы обучение заняло оптимальное время, в отличие от другой модели ‘bert-large-uncaused’.

BERT нуждается в особенном формате представления данных, для этого мы будем использовать маску, т.е. когда она обучается, то на входные данные использует маску. Мы используем attention mask и input ids: первая является необязательной, она состоит из последовательности нулей и единиц, где 1 - токены заголовка, а 0 - паддинги (необходим для работы BERT с заголовками разной длины). Input ids соотносят каждому токену номер из встроенного словаря.

Разделяем нашу выборку на обучающую, тренировочную и тестовую. Настраиваем параметры batch_size = 4 и epochs = 3. Авторы статьи [10] пишут, что не обязательно должно быть много слоев для обучения для данной модели, достаточно 2-4. Мы взяли оптимальное значение.

```
optimizer = tf.keras.optimizers.Adam(learning_rate=2e-5, epsilon=1e-08)
loss = tf.keras.losses.CategoricalCrossentropy()
metrics = [tf.keras.metrics.CategoricalAccuracy()]
```

Настраиваем функцию оптимизации, параметры оценки мы взяли из статьи авторов BERT по рекомендации [10]. Функция ошибки задана в виде категориальной перекрестной энтропии, основная метрика - accuracy.

```
Epoch 1/3
176/176 [==============================] - 166s 901ms/step - loss: 0.5259 - accuracy: 0.7320 - val_loss: 0.3367 - val_accuracy: 0.8628
Epoch 2/3
176/176 [==============================] - 147s 836ms/step - loss: 0.2822 - accuracy: 0.8924 - val_loss: 0.3135 - val_accuracy: 0.8780
Epoch 3/3
176/176 [==============================] - 142s 810ms/step - loss: 0.2118 - accuracy: 0.9221 - val_loss: 0.3208 - val_accuracy: 0.8808
```

Рисунок 14 - Эпохи прохождения сети по датасету

Изначально мы указывали batch_size = 32, но персональный компьютер не выдержал нагрузки и обучение не выполнилось. Пришлось уменьшить в 2 раза этот параметр, время обучения составило около 5 минут. Ниже представлен график, на котором виден прогресс обучения уже на втором прохождении модели по выборке.

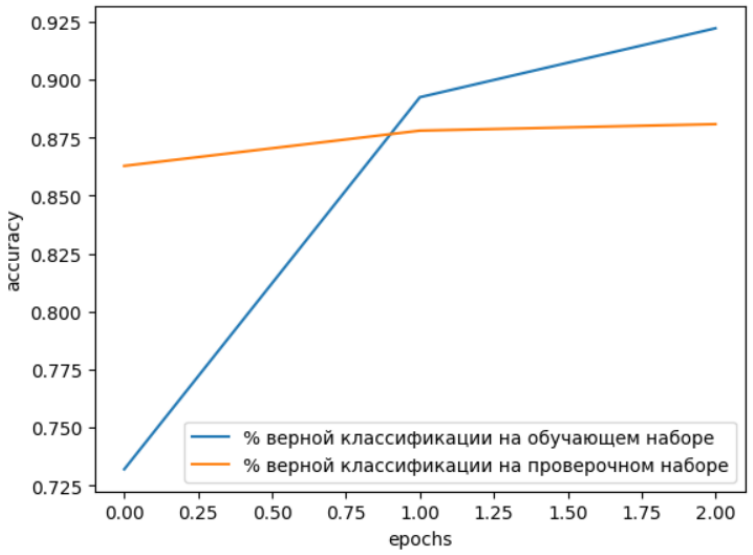


Рисунок 15 - График соотношения точности к эпохам на основе сравнения с обучающей выборкой и проверочной

Проверим качество обучения нашей модели на тестовой выборке, параметр accuracy составил 0,85684, что опять же ниже даже чем на тренировочной выборке.

2.4.5 RoBerta модель

Загружаем модель “Roberta-base” и его одноименный токенизатор. В принципе процесс реализации данной языковой модели схож с моделью BERT, разница между ними лишь в основном в производительности, считается что RoBerta точнее классифицирует, но делает это в 4-5 раз дольше чем BERT.

Мы практически не меняли алгоритм написания данной языковой модели, кроме названия с BERT на RoBERTa. Мы используем attention mask и input ids: первая является необязательной, она состоит из последовательности нулей и единиц, где 1 - токены заголовка, а 0 - паддинги (необходим для работы RoBERTa с заголовками разной длины). Input ids соотносят каждому токену номер из встроенного словаря.

Тем не менее, основные параметры мы указали следующие:

- max_length = 98 (длина токена)
- activation = ‘softmax’ (функция активации)
- epochs = 3 (в ходе работы подлежит регулированию)
- batch_size = 4

Функции оптимизации, ошибки и основная метрика качества остаются такими же. Мы получили средний результат в конце обучения, но время обучения действительно сильно увеличилось.

```
Epoch 1/3
147/147 [==============================] - 3186s 21s/step - loss: 0.6629 - categorical_accuracy:
0.6936 - val_loss: 0.4382 - val_categorical_accuracy: 0.7956
Epoch 2/3
147/147 [==============================] - 3081s 21s/step - loss: 0.3732 - categorical_accuracy:
0.8220 - val_loss: 0.5829 - val_categorical_accuracy: 0.7776
Epoch 3/3
147/147 [==============================] - 3098s 21s/step - loss: 0.2897 - categorical_accuracy:
0.8538 - val_loss: 0.4137 - val_categorical_accuracy: 0.8178
```

Рисунок 16 - Список эпох в RoBERTa

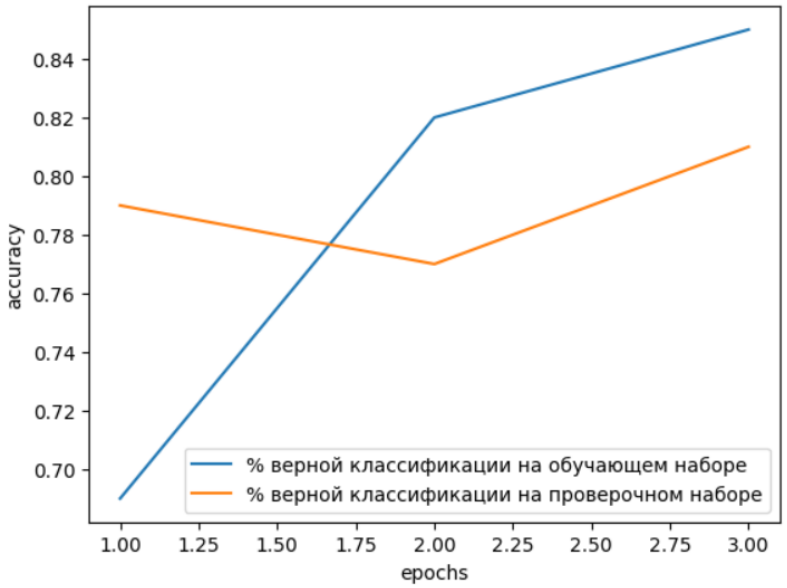


Рисунок 17 - График соотношения точности к эпохам на основе сравнения с обучающей выборкой и проверочной

Параметр accuracy в результате использования нашей модели RoBERTa на тестовой выборке составил 0,80394.

2.4.6 Выводы по результатам обучения моделей

Мы провели обучение на 5 разных языковых моделях, сравним их по характеристике точность и по времени обучения. Оба параметра являются важными и стоит учитывать их при выборе.

Таблица 1 - Сравнение accuracy классификации моделей

Модель	Точность (accuracy), %
Логистическая регрессия+BOW	51
Мультиномиальный наивный Байес+BOW	70
LSTM+word2vec	84
BERT	85
RoBERT	80

Таблица 2 - Сравнение времени обучения

Модель	Время обучения, секунды
Логистическая регрессия+BOW	2
Мультиномиальный наивный Байес+BOW	3
LSTM+word2vec	136
BERT	1423
RoBERT	3098

Больше всего нам подходит модель BERT, так как она обладает наибольшим показателем точности классификации. У нее есть недостатки: время обучения среднее, память обычного ноутбука сильно занимает, словарь слов не расширенный (по сравнению с RoBERT). Можно было выбрать LSTM модель, но у него огромный риск к переобучению, это было видно на графике, хоть и метрики качества были положительные, а время на обучение затрачено меньше чем на две остальных нейронных сети.

Изначально предполагалось, что наилучший результат будет у RoBERT, так как она имеет расширенный словарь, однако она не смогла показать результат лучше чем BERT, делаем предположение что это из-за сравнительно маленького датасета (538 строк).

В ходе обучения моделей были проведены эксперименты по подбору количества эпох и batch_size, были найдены оптимальные значения для классификации и исходя из возможностей используемой ЭВМ. На основе полученных метрик и графиков можно сказать, что классифицировать заголовок возможно.

3. Концепция стартап-проекта

Анализ фондового рынка всегда был привлекателен для исследователей, хочется уметь грамотно инвестировать свои средства, не теряя ничего. Разработано уже довольно много методов прогноза рыночных цен опционов. Существенный изъян данных техник заключается в том, что они рассчитаны исключительно с помощью случайных процессов, не включая семантику. Люди, занимающиеся техническим анализом, используют диаграммы и техники моделирования для определения тенденций в движениях цен и объемов продаж. Эти индивидуалисты полагаются на исторические данные, чтобы предсказать будущие результаты.

Практика прошлых лет показала, что технический прогноз бессмысленен без учета внешних различных событий, которые как раз напрямую влияют на ценовое движение. Мало предугадать рост цен, необходимо ориентироваться по ситуации в мире и принимать гибкие решения относительно опционов. Ниже на рисунке 17 представлен график сентимента, оказывающий влияние на опционную биржу, в очередной раз намекая как важно проводить анализ тональности финансовых событий.

Рисунок 18 - График сентимента [18]

В связи с этим было предложено анализировать внешние события, рассматривать их эмоциональную окраску. Анализ тональности относится к использованию обработки естественного языка, текстового анализа и вычислительной лингвистики для идентификации и извлечения субъективной информации в исходных материалах. Чем дольше вы применяете анализ рыночных настроений, тем более точным будет его результат, однако в некоторых случаях рынок может повести себя совершенно непредсказуемо и не оправдать ваши прогнозы. Поэтому нужно проявлять осмотрительность в вопросах трейдинга и инвестиций и помните, что каждое решение сопряжено с определенным риском. Вообще говоря, анализ тональности направлен на определение отношения говорящего или писателя к какой-то теме или на определение общей концептуальной направленности документа. Отношением может являться суждение или оценка, эмоциональное состояние или предполагаемый эмоциональный посыл. Основная задача при анализе тональности — классифицировать полярность данного текста в документе, предложении, любом другом уровне текста на позитивную, негативную или нейтральную.

В данном проекте был разработан наиболее оптимальный способ анализировать финансовые заголовки. Человек, воспользовавшийся этим технологическим решением, сможет принять решение относительно опциона, который упоминается в новости. При этом опираясь на объективный анализ искусственного интеллекта, который безусловно может ошибиться, но данный продукт не является инвестиционной рекомендацией и носит лишь информативный характер.

3.1. Описание продукта как результата НИР

В результате выполнения данной работы решается проблема пустой траты денежных ресурсов в ходе неправильного инвестирования, опираясь на математическое прогнозирование. Предлагаемое решение - программа, выполняющая анализ финансовых заголовках о компаниях, торгующих на бирже. Обеспечение выполняет роль информатора, который сообщает характер новости.

Далее инвестор принимает решение относительно опциона: покупать, держать или продавать. Таким образом, мы выполняем не технический анализ, который не основан на реальных событиях, а анализ тональности мировых новостей.

На данный момент проблема решается иностранными учеными, была разработана собственная модель, опирающаяся на собственный словарь слов. Она легко адаптируется к свежей сборке новостей, выполняет анализ за 10 секунд, имеет высокую точность - около 89%. Но главный недостаток - она не работает с русским языком, а только с английским. Мы строим собственную модель, адаптированную под русский язык и английский, ее главная задача - уметь хорошо обрабатывать текст. Используя хорошую предобработку данных, проведение лемматизации, удаление пустых слов, разбиение предложений на слова и приведение их в начальный вид, позволяет разработать высокоточное программное обеспечение.

Модель функционирует в онлайн режиме, скорость обработки и анализа данных составляет 3 минуты, общая точность определения правильной эмоциональной окраски - 86%. В дальнейшем можно адаптировать ее под собственное приложение или сайт, сделать автоматический сбор заголовков новостей.

Таким образом, было разработано программное обеспечение, которое выполняет следующие задачи:

1. Принимает входные данные в виде заголовков новостей о российских компаниях.
2. Определяет тональность заголовков, соотносит к трем видам: новость положительная, отрицательная или нейтральная.
3. Формирует ответ в виде таблицы с проанализированными заголовками, которые были загружены в начале.

3.2. Интеллектуальная собственность

Объект интеллектуальной области представляет собой программу для ЭВМ. В случае возникновения различных ситуаций, нужно обратиться к Гражданскому

кодексу РФ, часть 4, ст. 1261 и ст. 1262. Регулирование отношений происходит на законодательном уровне. Программное обеспечение, разработанное в рамках этой работы, не является изобретением согласно ст. 1350 п. 5 ГК РФ [1]. Программа состоит из алгоритма и исходного кода, признается патентоспособным только алгоритм. Исходный код охраняется авторским правом, оно защищает автора от копирования программного обеспечения, а патентное право защищает идею программы.

Анализируя вышенаписанное, было принято решение зарегистрировать исходный код и запатентовать алгоритм программного обеспечения. В случае судебного разбирательства, у нас будут сертификаты о наличии патента.

3.3. Объем и емкость рынка

Проведем анализ рынка:

1. Продукт: ПО, определяющее окраску заголовка новости о компании.
2. Конкурентные решения: приложения, определяющие тональность новости в финансовой сфере.
3. Целевая аудитория: Обычный потребитель, занимающийся инвестициями на любительском уровне.
4. География: Россия.

Оценка рынка произведена при помощи методов: снизу-вверх и сверху-вниз. Воспользуемся методом сверху-вниз:

TAM: Объем инвесторов в России составляет 12.7 миллионов людей. Из них 556 тысяч являются квалифицированными инвесторами, то есть им вряд ли понадобятся наши услуги [2]. Обозначим подписку на наше ПО - 200 рублей в месяц. Итого: TAM = 12144000 тысяч людей*200 рублей/месяц = 2 428 800 000 рублей.

SAM: Оценим, сколько человек из непрофессиональных инвесторов пользуется форумами или социальными сетями с финансовыми новостями. Согласно статистике [3], рекордное количество посетителей составляет 6 миллионов инвесторов. SAM = 200 рублей * 6000000 человек = 1 200 000 000 рублей.

SOM: Рекордное количество посетителей форумов не полностью заинтересовано в анализе финансового рынка, часть из них заходила на сайт за определенной информацией, не задерживаясь больше чем на одной вкладке или новости. Количество таких инвесторов составляет 40% [3], значит они вряд ли заинтересованы в тщательном анализе рынка опционов. Следовательно, SOM = 200 рублей * 5700000 (среднее количество инвесторов, пользующихся сообществом для инвесторов) человек * 0.6 = 684 000 000 рублей.

Проведем анализ рынка методом снизу-вверх:

TAM: На рынке существуют готовые сервисы для сентимент-анализа, общий объем достигает 128 478 000 000 рублей [4]. При этом только 10% сервисов способны анализировать заголовки новостей из финансовой сферы. TAM = 128478000000 рублей * 0.1 = 12 847 800 000 рублей.

SAM: Из оценки TAM мы не учли то, что только 7% сервисов [4] могут анализировать информацию на русском языке. Следовательно, SAM = 128478000000 * 0.07 = 899 346 000 рублей.

SOM: Остается учесть сервисы, недоступные в России, и сервисы с бесплатным использованием, они составляют [4] около 2% от SAM. Получается что SOM = 899346000 * 0.8 = 719 476 800 рублей.

Таким образом, наиболее оптимистичной оценка рынка является при методе сверху-вниз и предполагаемый объем рынка, который охватывает наш продукт составляет 684 000 000 рублей.

Таблица 3 - Сравнительный анализ рынка с помощью двух методов

Сверху-вниз, млн. руб.	Снизу-вверх, млн. руб.
TAM	
В России 12.7 инвесторов, из которых 556 тысяч являются квалифицированными. Подписка стои 200 рублей в месяц **12,144*200 руб. = 2 428, 8**	Объем рынка сервисов для сентимент-анализа составляет 128 478 000 000 рублей. Только 10% могут проводить анализ в финансовой сфере. **128, 478 * 0.07 руб. = 899, 346**
SAM	

Количество людей, которые пользуются форума или социальными сетями для инвесторов дости 6 миллионов. **200 руб. * 60 = 1 200**	Доля сервисов, способных работать с заголовками русском языке - 7% **128,478 * 0.07 = 899,346**
SOM	
Из посетителей форумов есть те, кто заходил за определенной информацией и не задерживался, таких людей 40% из всех пользователей **200 руб.* 5,7 = 684**	Количество сервисов, недоступных в России и бесплатных для использования, составляет 2% **899,346 * 0.8 = 719,4768**

3.4 Анализ современного состояния и перспектив развития отрасли

Сентимент-анализ набирает обороты каждый год, благодаря ему существует возможность быстро оценить настроение аудитории в какой-либо сфере, спрогнозировать мнение по определенному продукту, выявить паттерны настроения общества. Активно используется анализ тональности в бизнесе для подачи целевой рекламы. Сфера является актуальной в первую очередь из-за автоматизации оценки эмоциональной окраски.

В нашем случае сентимент-анализ проводится в финансовой сфере, существует множество готовых нейронных сетей, способных обучиться на определенных данных и классифицировать их на позитивные или негативные части. К примеру, нейронные сети FinBert обучены на специальных данных из новостей об акциях, способные обрабатывать кучу информации и делать определенные выводы об опционах [6].

В первую очередь, сентимент- анализ используется в маркетинге (выяснение мнения потребителя о продукте), политике (способность влиять на общественное мнение путем оценки), любой компании (с целью узнать отношение сотрудников к чему-либо), финансы (помощь в принятии решения относительно рынка). Анализ настроений позволяет трейдерам отслеживать рыночный спрос и прогнозировать потенциально прибыльные тренды. Он не всегда учитывает фундаментальные показатели того или иного проекта, но иногда эти факторы могут быть взаимосвязаны.

Анализ рыночных настроений – неотъемлемая часть многих торговых стратегий наряду с техническим и фундаментальным анализом. Он помогает инвесторам оценить всю доступную информацию перед принятием каких-либо решений. Так, например, анализ настроений поможет выяснить, оправдан ли FOMO (страх упущенной выгоды) участников рынка или просто вызван общественным мнением. В целом использование технических и фундаментальных анализов вместе с анализом сентимента позволяет:

- получить лучшее представление о краткосрочных и среднесрочных движениях цен;
- лучше контролировать свое эмоциональное состояние;
- обнаружить потенциальные возможности получения прибыли.

Таким образом, направление сентимент-анализа является развивающимся и перспективным в развитии.

3.5 Планируемая стоимость продукта

В расчете планируемой стоимости продукта для вычисления себестоимости используется затратный метод ценообразования. В данную статью расходов следует включить стоимость компьютеров и программного обеспечения в виде амортизационных отчислений. Поскольку для проведения исследований специальное дорогостоящее оборудование не приобреталось, при расчете затрат учитывается только амортизация.

Первоначальная стоимость ПК исполнителя, используемого для проведения исследований, составляет 49000 рублей. Срок полезного использования данной машины – 3 года, из которых 3 месяца машина использовалась для написания проекта.

Норма амортизации: Ан = 1/ n ∗ 100% = 1/3 × 100% = 33,33%

Годовые амортизационные отчисления: Аг = 49000 × 0,33 = 16170 рублей

Ежемесячные амортизационные отчисления: Ам = 16170/ 3 = 5390 рублей

Итоговая сумма амортизации основных средств: А = 5390*3= 16170 рублей

Таким образом, затраты на амортизацию ПК составляют 16170 рублей.

Проведем расчет амортизации за сервер для данных:

Норма амортизации: Ан = 1/ n ∗ 100% = 1/3 × 100% = 33,33%

Годовые амортизационные отчисления: Аг = 50000 × 0,33 = 16500 рублей

Ежемесячные амортизационные отчисления: Ам = 16500/ 3 = 5500 рублей

Итоговая сумма амортизации основных средств: А = 5500*3= 16500 рублей

Таблица 4 - Расчет затрат по статье «Спецоборудование для научных работ»

№ п/п	Наименование оборудования	Кол-во единиц оборудования	Стоимость единицы оборудования, тыс.руб.	Общая стоимость оборудования, тыс.руб.
1.	Ультрабук Honor Magicbook	1	49	49
2.	Дистрибутив языка программирования Python – Anaconda Individual Edition		0 (распространяется по модифицированной лицензии BSD свободного ПО [4])	0
3.	Сервис Google Collaboratory		0	0
4.	Сервер для хранения данных		50	50

Рассчитаем заработную плату сотрудников, нам для выполнения понадобятся программист и аналитик машинного обучения. Рассмотрим количество рабочих дней:

Таблица 5 - Баланс рабочего времени

Показатели рабочего времени	Аналитик	Программист
Календарное число дней	365	365
Количество нерабочих дней	118	118

Потери рабочего времени - отпуск - невыходы по болезни	24	24
Действительный годовой фонд рабочего времени	223	223

Месячный должностной оклад работника:

$$З_м = З_б \cdot \left(1 + k_{пр} + k_д\right) \cdot k_р,$$

где $З_б$ – базовый оклад, руб.;

$k_{пр}$ – премиальный коэффициент;

$k_д$ – коэффициент доплат и надбавок;

$k_р$ – районный коэффициент, равный 30% (для Томска).

Исполнителями темы выступают младший аналитик машинного обучения и начинающий программист. Оклад младшего аналитика составляет 37700 рублей, оклад программиста – 19200 рублей [5]. Аналитик не работает полный день, поэтому возьмем понижающий коэффициент как половину ставки - 19350 рублей в месяц.

Месячный должностной оклад программиста:

$$З_м = 19200 \cdot (1 + 0 + 0) \cdot 1,3 = 24960 \text{ рублей}$$

Месячный должностной оклад аналитика:

$$З_м = 19350 \cdot (1 + 0 + 0) \cdot 1,3 = 25155 \text{ рублей}$$

Расчет заработной платы за день для программиста:

$$З_{дн} = \frac{З_м \cdot М}{F_д} = \frac{З_м * \left(1 + k_{пр} + k_д\right) \cdot k_р * 10,4}{223} = \frac{24960 * 1,3 * 10,4}{223} = 1513,27 \text{ рублей}$$

Расчет заработной платы платы за день для аналитика:

$$З_{дн} = \frac{З_м \cdot М}{F_д} = \frac{З_м * \left(1 + k_{пр} + k_д\right) \cdot k_р * 10,4}{223} = \frac{25155 * 1,3 * 10,4}{223} = 1525,09 \text{ рублей}$$

Основная заработная плата программиста проекта рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot Т_{раб} = 1513,27 * 45 = 68097,15 \text{ рублей}$$

Основная заработная плата аналитика проекта:

$$З_{осн} = З_{дн} \cdot Т_{раб} = 1525,09*29=44227,61 \text{ рублей}$$

Таблица 6 - Расчет основной заработной платы

Исполнители	$З_б$, руб.	$k_{пр}$	$k_д$	$k_р$	$З_м$, руб	$З_{дн}$, руб.	$Т_р$, раб. дн.	$З_{осн}$, руб.
Аналитик	37700	0	0	1,3	25155	1525, 09	29	44227,61
Программист	19200	0	0	1,3	24960	1513,27	45	68097,15

Рассчитаем отчисления во внебюджетные фонды, при условии, что коэффициент отчислений $k_{внеб} = 30,2\%$.

$$C_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп})$$

$С_{внеб} = 0.302 \cdot (44227,61 + 68097,15) = 33922,0775$ руб.

В данную статью расходов при выполнении проекта отнесем использование Internet. Оплата подключения к сети Internet производится один раз в месяц в размере 350 рублей. Проект длится 2 месяца, значит суммарно будет потрачено 2·350 = 700 рублей.

В прямые затраты также следует включить затраты на электроэнергию, потребляемую оборудованием. Стоимость 1 кВт электроэнергии составляет 5,8 руб., мощность одного используемого ноутбука 33 Вт/ч, коэффициент использования мощности – 0,8 [6], суммарное количество часов работы ноутбука (29 + 45) · 8 = 592. Суммарная потребляемая энергия составляет 33 · 0,8 · 592 = 15628,8 Вт, её стоимость равна 15,628 · 5,8 = 90,6424 руб.

Накладные расходы проекта составляют 70% от суммы основной и дополнительной зарплат сотрудников. Поскольку дополнительная зарплата сотрудникам не предусмотрена, накладные расходы будут составлять

$С_{накл} = 0.7 \cdot 112324,76 = 78627,332$ руб.

Таблица 7 - Бюджет проекта

Наименование статьи	**Сумма, руб.**

1. Затраты на амортизацию	32610
2. Затраты по основной заработной плате исполнителей темы	112324,76
3. Отчисления во внебюджетные фонды	33922,0775
4. Контрагентские расходы	700
5. Накладные расходы	78627,332
6. Прочие прямые затраты	90,6424
7. Бюджет затрат НТИ	241834,812

Таким образом, на реализацию данного проекта потребуется 241834,812 рублей.

3.6 Конкурентные преимущества создаваемого продукта

При проведении конкурентного анализа были выбраны наиболее близкие по концепции проекты, способные придавать эмоциональную окраску новостям. Аналогичный сервис не удалось найти, только прототипы, не попавшие на рынок, поэтому будем рассматривать те варианты, которые активно используются инвесторами. Зачастую это обыкновенные сайты-форумы, на которых разные опытные люди ведут блоги, в которых они описывают новость и делают соответствующие выводы об опционе. Среди них наиболее популярны такие сервисы, как: “Smart-lab”, “Investing”, “Финам”. Сведения об этих продуктах были взяты из открытых источников, а именно непосредственно на их сайтах [14].

Для сравнения продуктов мы выделили такие важные критерии, как: масштабность, оперативность, объективность, доступность, отсутствие визуального шума. Масштабность важна для пользователя потому что она позволяет проанализировать сразу несколько новостей или любой заголовок, найденный в сети. Оперативность высоко ценится среди пользователей, позволяет не ждать слишком долго анализ рынка. Объективность, она же непредвзятость, мнение анализатора не должно зависеть от отношения к

компании, которая имеет свой опцион, иначе новость проанализирована некорректно. Продукт должен быть доступен в любой момент для инвестора, потому что на рынке всегда что-то происходит, важно успеть среагировать на действия опционов. И наконец, отсутствие визуального шума или лаконичность - инвестору важно быстро получить анализ по опциону, не тратя много времени на кучу информации, а визуальный шум всегда отвлекает от важного и не дает принимать быстро решение. Также можно долго анализировать самостоятельно разные графики или параметры опциона, но тогда необходимость в стороннем анализе отпадает.

Проведенный конкурентный анализ представлен в таблице 1.3.

Таблица 8 - Конкурентный анализ

Решение / Критерий	Разрабатываемое решение	Финам	Smart-lab	Investing
Масштабность	+	-	-	-
Оперативность	+	+	-	+
Объективность	+	-	-	-
Доступность	+	+	+	-
Отсутствие визуального шума	+	-	-	-

Таким образом, согласно нашим критериям главным конкурентом является сервис “Финам”, главными преимуществами нашего продукта являются масштабность (возможность проанализировать любой заголовок) и объективность (человек всегда субъективен, в отличие от искусственного интеллекта).

3.7 Бизнес-модель проекта. Производственный план и план продаж.

Приведем описание бизнес-модели нашего проекта с помощью схемы. Для этого прибегнем к модели А. Остервальдера, которая схематично описывает с помощью 9 блоков различные бизнес-процессы проекта. Мы рассмотрим потенциальных партнеров, выделим ключевые виды деятельности и ресурсы

нашего проекта. Выделим предложение на рынке, способы взаимодействия с клиентами, способы привлечения клиентов, а также потенциальную аудиторию продукта.

Бизнес-модель нам необходима для выделения логики процесса создания ценности, а также нахождения взаимосвязи блоков, которые представляют собой 4 основных сферы бизнеса: продукт, взаимодействие с потребителем, инфраструктура, финансовая эффективность.

В таблице 9 описана бизнес-модель по А. Остервальдеру.

Таблица 9 - Бизнес-модель по А. Остервальдеру

<table>
<tr><td rowspan="2">Ключевые партнеры
Инвесторы, занимающиеся непрофессионально покупкой-продажей опционов на бирже</td><td>Ключевые виды деятельности
Разработка программного обеспечения
Техническая поддержка</td><td rowspan="2">Ценностные предложения
Информация о возможном поведении опционов на бирже</td><td>Взаимоотношения с клиентами
Персональная техническая поддержка
Автоматический сервис</td><td rowspan="2">Потребительские сегменты
Люди, принимающие активное участие в инвестициях на любительском уровне</td></tr>
<tr><td>Ключевые ресурсы
Финансы (з/п сотрудникам)
Средства для размещения базы данных на сервере</td><td>Каналы сбыта
Основной канал сбыта-информационный: реклама, использование таргета, социальные сети</td></tr>
<tr><td colspan="3">Структура издержек
Постоянные: налоги, заработная плата сотрудникам на разработку программного обеспечения, оплата за размещение базы данных на сервере
Переменные: заработная плата сотрудникам для поддержания программного обеспечения</td><td colspan="2">Потоки поступления доходов
Продажа подписки на использование сервиса</td></tr>
</table>

Опишем также производственный план проекта, в котором покажем основные задачи при выпуске продукта. Здесь мы учитываем организационные моменты, технический анализ и реализацию проекта. На выполнение задач дается максимум месяц в рабочем режиме, зона ответственности одинаковая, т.е. на задачами работают все причастные к проекту люди.

Таблица 10 - Производственный план

Наименование задачи	Июль 2023 г.	Август 2023 г.	Сентябрь 2023 г.
Поиск работников			
Разработка структуры проекта			
Закуп оборудования			
Создание выборки для моделей			
Предобработка выборки			
Реализация нескольких языковых моделей			
Анализ полученных результатов			
Тестирование лучшей модели			
Запуск проекта			

3.8 Стратегия продвижения продукта на рынок

Целью продвижения продукта на рынок является использовать общественные каналы, формируя бесплатное распространение сервиса среди пользователей. Для этого можно прибегнуть к активным дискуссиям на форумах среди инвесторов, в ходе которых можно порекомендовать использовать данный сервис. Приоритетные точки касания представлены в виде активного обсуждения, использования своего блога, переписки с инвесторами. Можно прибегнуть к использованию таргетированной рекламы на известных поисковиках. То есть пользователь будет искать инвестиционный форум с профессиональными инвесторами, а ему будет попадаться реклама нашего сервиса, который выигрышнее выглядит. Также в целях продвижения будет личное посещение форумов инвесторов и различных собраний, на которых можно подготовить доклад на тему “Почему использование искусственного интеллекта лучше чем

использование рекомендаций живых людей при инвестировании” и в конце предложить использование данного сервиса. В случае использования таргета, на баннере можно предложить использование пробного периода подписки, чтобы пользователь понял целесообразность приобретения полной подписки на сервис.

Определим путь клиента при продвижении:

1. Увидел рекламный баннер с продуктом
2. Перешел по ссылке/сохранил название
3. Ознакомился с использованием продукта
4. Приобрел подписку

Первый пункт продвижения появляется как при личном продвижении, так и при использовании таргетированной рекламы или в ходе выступления на мероприятиях. Второй пункт чередуется в зависимости от формата продвижения, описанных выше. Третий пункт подразумевает использование пробного периода подписки. И наконец, четвертый пункт не является конечным, поскольку подписку можно продлить на любое время.

3.9 Вывод по разделу “Концепция стартап-проекта”

В данном разделе была разработана идея для реализации данной ВКР в форме стартап-проекта. Для ее успешной реализации были проведены мероприятия по защите интеллектуальной собственности - патентование алгоритма, оценен объем рынка продукта - 684 млн. руб., рассмотрены перспективы развития. Подсчитана примерная себестоимость реализации - 241834,812 руб. за 3 месяца проведения работы. Выполнен анализ конкурентов, найдены преимущества и недостатки, главное отличие нашей работы - взаимодействие с русским языком. Построена бизнес-модель по Остервальдеру, в которой рассмотрены ценностное предложение, каналы сбыта - в основном реклама, основной заработок - подписка на сервис, потенциальная аудитория - непрофессиональные инвесторы, ключевые ресурсы. Расписали стратегию продвижения и реализации продукта на рынке, примерный путь потенциального покупателя состоит из 4 шагов.

4. Социальная ответственность

4.1 Введение

Объектом разработки данной ВКР является программное обеспечение для сентимент-анализа заголовков новостей о компаниях, чьи опционы существуют на бирже. Методы машинного обучения, использованные при разработке классификаторов, предоставляют возможность анализировать эмоциональную составляющую исходных данных. Алгоритм предобработки и очистке данных позволяет повысить точность обучения интеллектуальной системы.

Проект выполняется на персональном компьютере (ПК), поэтому в данном разделе проводится анализ опасных и вредных факторов при работе с ПК, влияния этих факторов на окружающую среду и мероприятий по ее защите.

Предметом исследования является рабочая зона разработчика, включая компьютерный стол, ПК, клавиатуру, компьютерную мышь и стул. Работы выполняются в компьютерном классе 427А 10 корпуса ТПУ.

4.2 Правовые и организационные вопросы обеспечения безопасности

Процесс разработки программного обеспечения происходит за компьютерным столом. Рабочее место должно удовлетворять требованиям ГОСТ 12.2.032-78 «Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя» [16] РД 153-34.0-03.298-2001 «Типовая конструкция по охране труда для пользователей персональными электронно-вычислительными машинами (ПЭВМ) в электроэнергетике» . Выделяют следующие основные требования к рабочему месту:

1. Рабочее место должно быть организовано с учетом эргономических требований:

а) рабочий стол может быть любой конструкции, отвечающей современным требованиям эргономики и позволяющей удобно разместить на рабочей

поверхности оборудование с учетом его количества, размеров и характера выполняемой работы;

б) рабочие места с ПК по отношению к световым проемам должны располагаться так, чтобы естественный свет падал сбоку, желательно слева. Схемы размещения рабочих мест с ПК должны учитывать расстояние 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 составляет 15° вниз от горизонтали).

2. Конструкция рабочей мебели (рабочий стол, кресло, подставка для ног) должна обеспечивать возможность индивидуальной регулировки соответственно росту пользователя и создавать удобную позу для работы. Также вокруг ЭВМ должно быть обеспечено свободное пространство не менее 60–120 см;

Требования к нормам труда (продолжительность рабочего дня, перерывы в течение рабочего дня, перерывы на обед) регламентируются ТК РФ «Рабочее время» . При 8-ми часовой рабочей смене регламентированные перерывы следует устанавливать через 1,5–2,0 часа от начала работы и через такой же промежуток после обеденного перерыва продолжительностью 20 минут каждый или продолжительность 15 минут через каждый час работы. Суммарное время регламентированных перерывов составляет от 30 до 120 минут в соответствии с категорией работ (ст. 108 ТК РФ). Согласно классификации видов трудовой деятельности с персональным компьютером (ТОИ Р-45-084-01 Типовая инструкция по охране труда при работе на

персональном компьютере), работу разработчика следует отнести к группе В, которая предполагает работу в режиме диалога с компьютером [17].

4.3 Производственная безопасность

В ходе создания программного обеспечения разработчики подвергаются воздействию различных вредных и опасных факторов, которые представлены в таблице 1.1. В таблице также находятся соответствующие нормативные документы и этапы работ, во время выполнения которых разработчики могут столкнуться с их влиянием.

Таблица 11 – Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
Отклонение показателей микроклимата	ГОСТ 30494-2011 «Здания жилые и общественные. Параметры микроклимата в помещениях»
Недостаточная освещенность рабочей зоны	СП 52.13330.2016 «Естественное и искусственное освещение» [5]
Повышенная световая и цветовая контрастность	СП 52.13330.2016 «Естественное и искусственное освещение» [5]
Повышенный уровень шума на рабочем месте	ГОСТ 12.1.003-83 «Система стандартов безопасности труда. Шум. Общие требования безопасности»
Повышенный уровень статического электричества	ГОСТ Р 53734.1-2014 «Электростатические явления» [7]
Опасность поражения электрическим током	ГОСТ Р 58698-2019 «Защита от поражения электрическим током» [9]

Судя по данной таблице, на разработчиков программного обеспечения воздействуют только физические и психологические факторы, а химические и биологические факторы отсутствуют.

4.3.1. Отклонение показателей микроклимата

Существенное влияние на здоровье работника оказывает отклонение показателей микроклимата на рабочем месте. При повышенной скорости ветра и низкой температуре человек может столкнуться с переохлаждением

организма путем усиления теплообмена и процесса теплоотдачи при испарении пота. Недостаточная влажность в помещении ведет к интенсивному испарению влаги у слизистых оболочек, что приводит к пересыханию, растрескиванию и далее к появлению болезнетворных бактерий. Так как в ходе разработке программного обеспечения используются персональные компьютеры, они прямым образом влияют на микроклимат путем снижения важности и повышению температуры на рабочем месте в помещении.

Требования к микроклимату производственных помещений регламентируются ГОСТ 30494-2011 «Здания жилые и общественные. Параметры микроклимата в помещениях». Санитарные нормы регулируют оптимальные и допустимые значения показателей в рабочей зоне, соответствующие физиологическим потребностям организма человека, для создания комфортных и безопасных условий труда.

По энергозатратам работу, выполняемую разработчиками программного обеспечения, можно отнести к категории Iа (производится сидя, сопровождается незначительными физическими усилиями). В таблицах 1.2 и 1.3 представлены оптимальные и допустимые значения показателей микроклимата на рабочем месте для данной категории.

Таблица 12 - Оптимальные величины показателей микроклимата

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

Таблица 13 - Допустимые величины показателей микроклимата на рабочем месте

Период года	Температура воздуха, °С	Температура поверхнос	Относительная влажность	Скорость движения воздуха, м/с

	диапазон ниже оптимальных величин	диапазон выше оптимальных величин	тей , °С	воздуха, %	для диапазона температур воздуха ниже оптимальных величин, не более	для диапазона температур воздуха выше оптимальных величин, не более
Холодный	20,0 - 21,9	24,1 - 25,0	19,0 - 26,0	15 - 75	0,1	0,1
Теплый	21,0 - 22,9	25,1 - 28,0	20,0 - 29,0	15 - 75	0,1	0,1

В случае если поддерживать допустимые нормативные величины локального микроклимата не представляется возможным, нужно проводить различные такие мероприятия по защите работников от охлаждения и перегревания, как: проветривание, использование разных систем кондиционирования воздуха, регламентирование периодов работы в неблагоприятном микроклимате и отдых в помещении, которые нормализует тепловое состояние работника и т.д.

4.3.2. Недостаточная освещенность рабочей зоны

Недостаточную освещенность рабочего места относят к вредным производственным факторам, она приводит к повышенной утомляемости и снижает работоспособность разработчика. Если продолжительно работать в условиях недостаточной освещенности, можно ухудшить зрение.

Нормы естественного, искусственного и совместного освещения регламентируются СП 52.13330.2016 «Естественное и искусственное освещение». Разработка программного обеспечения относится к категории работ высокой точности – Б (наименьший или эквивалентный объект различения 0,30 94 – 0,50 мм), подразряд 1 (относительная продолжительность зрительной работы при направлении зрения на рабочую поверхность не менее 70%). В таблице 1.4 представлены требования к освещению рабочего помещения для разряда Б1.

Таблица 14 - Требования к освещению рабочего помещения

Искусственное освещение				Естественное освещение	
Освещенность на рабочей поверхности от системы общего освещения, лк	Цилиндрическая освещенность	Объединенный показатель дискомфорта, не более	Коэффициент пульсации освещенности, Кп, %, не более	Коэффициент естественной освещенности, %, при	
				верхнем или комбинированном	боковом
300	100	21	15	3	1

Присутствие яркого света в зоне периферийного зрения сильно увеличивает напряжение глаз. Для того чтобы снизить влияние вредного фактора недостаточной освещенности, нужно чтобы уровень естественного освещения приблизительно соответствовал яркости дисплея. Проблема недостаточной освещенности решается с помощью добавления искусственных источников света, увеличения количества окон, расширения световых проемов.

4.3.3. Повышенная световая и цветовая контрастность

Отклонение от нормы светового и цветового контраста на рабочем месте приводит к быстрой утомляемости и снижению работоспособности разработчика. В случае длительности воздействия этого вредного фактора ухудшается зрение. Нормы светового и цветового контраста регламентируются СП 52.13330.2016 «Естественное и искусственное освещение». Для работы за компьютером (категория работ Б1) нормы контраста представлены в таблице 1.5.

Таблица 15 - Нормы освещения рабочего помещения

Характеристика зрительной работы	Контраст объекта с фоном	Характеристика фона
Высокой точности	Малый	Средний
	Средний	Темный

Для решения проблемы светового и цветового контраста необходимо отрегулировать уровень естественной и искусственной освещенности рабочего помещения или заменить текущее оборудование (мониторы) на более качественные, которые позволят сгладить контраст.

4.3.4. Повышенный уровень шума на рабочем месте

Повышенный уровень шума на рабочем месте вызывает психологический стресс, который снижает продуктивность, концентрацию, внимание, работник быстро устает. В данной ситуации, на повышение уровня шума влияет фон, появляющийся из-за работы персональных компьютеров, а также при наличии систем кондиционирования и вентиляции воздуха.

Предельно допустимые показатели уровня звука, звукового давления регламентируются ГОСТ 12.1.003-83 «Система стандартов безопасности труда. Шум. Общие требования безопасности». Уровень шума на рабочем месте инженера программиста, который создает программу на ЭВМ, принтера, кондиционера, не должен превышать 50 дБА.

Для снижения уровня шума в рабочей зоне можно: заменить шумное техническое оборудование на менее шумное, использовать специальные звукопоглощающие экраны, регулярно проводить технический осмотр и обслуживание оборудования, так как загрязнение увеличивает производимый шум.

4.3.5. Повышенный уровень статического электричества

Статическое электричество является опасным производственным фактором, проявление которого может нанести вред здоровью человека (ожоги) или привести пожару и другим чрезвычайным ситуациям. При работе за компьютером статический заряд может накапливаться, если нет хорошего контакта с землей или влажность/ионизация воздуха превышает допустимые нормы. Статический разряд в производственных помещениях 97

рассматриваемого типа при условии соответствии нормам микроклимата и организации работ при воздействии на человека вызывает дискомфорт.

Допустимые показатели уровня статического электричества на производстве регламентируются ГОСТ Р 53734.1-2014 «Электростатические явления». В таблице 1.7 представлены уровни восприятия электростатического заряда человеком.

Таблица 16 - Уровни восприятия людьми электростатического заряда и ответной реакции при емкости тела в 200 пФ

Энергия разряда, мДж	Реакция	Потенциал тела, В
0,1	Разряд ощутим	1000
0,9	Четко ощутим	3000
6,4	Неприятный шок	8000

Чтобы уменьшить накопление статического заряда при работе за персональным компьютером, необходимо соблюдать норму влажности воздуха и поддерживать чистоту помещения, так как пыль имеет свойства диэлектрика.

4.3.7. Опасность поражения электрическим током

Электробезопасность подразумевает под собой систему технических и организационных мероприятий, направленных на защиту работников от опасного влияния и воздействия электрического тока, статического электричества и электромагнитного поля. Значения вышеперечисленных факторов регулируются ГОСТ Р 58698-2019. В таблице 1.9 представлены нормы напряжения прикосновения для реагирования.

Таблица 17 - Пороги напряжения прикосновения для реагирования

Характер реагирования	Пороги напряжения, В
Реакция испуга	2 (переменный ток)

	8 (постоянный ток)
Мышечная реакция	20 (переменный ток)
	40 (постоянный ток)

Для того чтобы избежать поражение электрическим током, нужно принять следующие меры предосторожности: размещение опасных для жизни и здоровья человека участков электропроводов и приборов вне зоны досягаемости телом; автоматическое отключение питания (защитное устройство, которое будет отключать систему, питающую электрическое оборудование в случае замыкания); использование защитных ограждений или оболочек; ограничение напряжения или питание должно осуществляться от безопасного источника питания.

Защита от поражения электрическим током может осуществляться посредством системы безопасного сверхнизкого напряжения (БСНН) и защитного сверхнизкого напряжения (ЗСНН).

4.4 Экологическая безопасность

Использование программного обеспечения не оказывает негативного влияния на окружающую среду. Однако использование самого компьютера оказывает влияние на окружающую среду.

Согласно ГОСТ Р 56397-2015 «Техническая экспертиза работоспособности радиоэлектронной аппаратуры, оборудования информационных технологий, электрических машин и приборов. Общие требования» пункт 5.8.1, после проведения технической экспертизы если оборудование не подлежит ремонту, то оно признается неработоспособным и рекомендуется к списанию (замене); в случае деградационного отказа оборудования и нецелесообразности его ремонта и модернизации даются рекомендации о необходимости его списания и утилизации.

Опираясь на «Методику проведения работ по комплексной утилизации вторичных драгоценных металлов из отработанных средств вычислительной 90 техники», утвержденной Государственным Комитетом РФ по телекоммуникациям от 19 октября 1999 г., п. 3.1.3. «Технология разборки универсальных ЭВМ» выделено 4 этапа разборки и подготовки к утилизации внутренних частей ПК. В результате выполнения этапов формируется партия сырья, включающая сортировку электронного лома по типу, проведение расчета количества ячеек, соединителей, серебросодержащих кабельных изделий, ячеек и типовых элементов замены. Элементы, содержащие драгоценные металлы и (или) партии черных и цветных металлов и сплавов (медь, сталь, никель, латунь, бронза, алюминий, дюралюминий) направляются на переработку на заводы ВДМ, полупроводниковые приборы (диоды, транзисторы), микросхемы в металлических и металлокерамических корпусах, а также конденсаторы в металлических корпусах демонтируются с плат и сортируются по типу, интегральные микросхемы в пластмассовых корпусах (серии 155, 551 и пр.) демонтируются и собираются отдельно, керамические конденсаторы типа КМ и резисторы после демонтажа также собираются отдельно. На рабочем месте программиста используются люминесцентные лампы ЛБ40. Согласно п. 2.1 ГОСТ 12.3.031-83 «Работы со ртутью. Требования безопасности», все ртутьсодержащие отходы и вышедшие из строя приборы, содержащие ртуть, подлежат сбору и возврату для последующей регенерации ртути в специализированных организациях.

4.5. Безопасность в чрезвычайных ситуациях

Причинами возникновения пожара при работе с компьютером (по ГОСТу 12.1.044-89 «Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения») может служить короткое замыкание проводки, вследствие неисправности прибора, сильный перегрев оборудования в результате его использования в режиме повышенной нагрузки. Для

предотвращения пожара, необходимо проводить своевременную диагностику оборудования и электрической проводки, обеспечить наличие средств пожаротушения в рабочем помещении, готовых к эксплуатации.

Здание, внутри которого помещение с ПК, тоже должно отвечать требованиям пожарной безопасности. Опираясь на ГОСТ 12.1.004-91 «ССБТ. Пожарная безопасность. Общие требования», для этого в нем должны быть пожарная сигнализация, план эвакуации, огнетушители с проверенным клеймом, знаки с указанием направления к эвакуационному выходу. На основании Федерального закона от 22.07.2008 N 123-ФЗ (ред. от 30.04.2021) "Технический регламент о требованиях пожарной безопасности" помещения должны быть оборудованы следующими средствами пожаротушения: огнетушитель ручной углекислотный ОУ-5, пожарный кран с рукавом, также каждое помещение оборудовано системой противопожарной сигнализации. Огнетушащие вещества должны обеспечивать тушение пожара поверхностным или объемным способом их подачи с характеристиками подачи огнетушащих веществ в соответствии с тактикой тушения пожара (согласно СП 484.1311500.2020 «Системы противопожарной защиты. Системы пожарной сигнализации и автоматизация систем противопожарной защиты. Нормы и правила проектирования»). Технические средства автоматических установок пожарной сигнализации должны быть обеспечены бесперебойным электропитанием на время выполнения ими своих функций.

4.6. Вывод по разделу “Социальная ответственность”

Значение всех производственных факторов на изучаемом рабочем месте разработчика соответствует нормам, которые были продемонстрированы в данном разделе. Не рассматривался фактор психофизиологического воздействия на организм человек, но для этого достаточно соблюдать меры из МР 2.2.9.2311 – 07 «Профилактика стрессового состояния работников при различных видах профессиональной

деятельности. Категория помещения по электробезопасности (согласно ПУЭ) соответствует первому классу – «помещения без повышенной опасности».

Придерживаясь правил по охране труда при эксплуатации электроустановок, персонал должен обладать I группой допуска по электробезопасности. Присвоение группы I по электробезопасности производится путем проведения инструктажа, который завершается проверкой знаний в форме устного опроса и проверкой приобретенных навыков безопасных способов работы или оказания первой помощи при поражении электрическим током. Категория тяжести труда в офисном помещении по СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" относится к категории Iб (работы, производимые сидя, стоя или связанные с ходьбой и сопровождающиеся физическим напряжением).

Помещение офиса категории помещения группы А, возможный класс пожара А и Е. Характеристика веществ и материалов, находящихся в помещении: горючие газы, легковоспламеняющиеся жидкости с температурой вспышки не более 28 °С в таком количестве, что могут образовывать взрывоопасные парогазовоздушные смеси, при воспламенении которых развивается расчетное избыточное давление взрыва в помещении, превышающее 5 кПа, и (или) вещества и материалы, способные взрываться и гореть при взаимодействии с водой, кислородом воздуха или друг с другом, в таком количестве, что расчетное избыточное давление взрыва в помещении превышает 5 кПа. Рассмотренный объект, оказывающий незначительное негативное воздействие на окружающую среду, относится к объектам III категории.

Заключение

В рамках данной выпускной квалификационной работы было разработано программное обеспечение по сентимент-анализу заголовков новостей компаний, торгующих на бирже. Для этого был реализован алгоритм предварительной обработки данных, в котором выборка была подготовлена к дальнейшему обучению. Используя стандартные модели классификации: “Bag of words” и “word2vec”, была проведена векторизация.

В качестве главного компонента программного обеспечения были построены различные классификаторы и языковые модели, такие как: классификатор логистической регрессии, мультиномиальный наивный Байес, нейронная сеть LSTM, BERT и RoBERT. В качестве оценки качества обучения моделей была использована метрика точности. С помощью этого параметра была выбрана наиболее подходящая модель под данную задачу. Мы получили неудовлетворительные результаты обучения классификатора логистической регрессии, он показал плохую матрицу ошибок, наихудший показатель полноты (recall) определения положительной новости, однако время обучения было наименьшим среди всех моделей, общая точность составляет 51%. Мы делаем вывод, что модель линейной регрессии неудовлетворительно работает с мультиномиальной классификацией и больше адаптирована под два вида классов. Классификатор Наивного Байеса продемонстрировал среднюю общую точность - 70%, время обучения быстрое, чуть дольше чем у логистической регрессии. Эта модель хорошо работает с тремя классами, матрица ошибок удовлетворительная. Далее мы перешли на обучение языковых моделей в виде нейронных сетей, в частности рекуррентная сеть LSTM, она отличилась быстрой обучаемостью и вторым показателем точности среди всех моделей. Но она уступает BERT и RoBERT маленьким объемом словаря (30 тыс. слов) для проведения векторизации выборки, что затрудняет адаптацию модели под большую выборку данных. Оптимальной моделью сентимент-анализа оказалась языковая модель BERT,

она продемонстрировала около 85% показателя определения правильных ответов, имела среднее время обучения - около 5 минут, использует свой словарь из 100 тыс. слов.

Список использованных источников

1. Лысенко, В. Д. Анализ тональности текста для прогнозирования цен на фондовом рынке / В. Д. Лысенко. — Текст : непосредственный // Молодой ученый. — 2018. — № 22 (208). — С. 420-423. — URL: https://moluch.ru/archive/208/51025/ (дата обращения: 24.05.2023).
2. Проблема классификации текстов и дифференцирующие признаки / И.В. Поляков, Т.В. Соколова, А.А. Чеповский, А.М. Чеповский // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. №. 2. С. 55–63.
3. Гречачин В.А. К вопросу о токенизации текста // Международный научно-исследовательский журнал. 2016. №. 6 (48). С. 25–27.
4. Wallach H.M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning. 2006. Pp. 977–984.
5. Rong X. word2vec parameter learning explained // arXiv preprint arXiv:1411.2738. 2014.
6. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey // Ain Shams engineering journal. 2014. Vol. 5(4). Pp. 1093–1113.
7. Проблема классификации текстов и дифференцирующие признаки / И.В. Поляков, Т.В. Соколова, А.А. Чеповский, А.М. Чеповский // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. №. 2. С. 55–63.
8. Tarasov D. S. Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews // Proceedings of the 21st international conference on computational linguistics dialog. 2015. Vol. 2. Pp. 53–64.

9. Feature selection for text classification with Naïve Bayes / J. Chen, H. Huang, S. Tian, Y. Qu //Expert Systems with Applications. 2009. Vol. 36(3). Pp. 5432–5435.
10. Understanding LSTM Networks. URL: http://colah.github.io/ posts/2015-08-Understanding-LSTMs (дата обращения: 21.05.2023).
11. Horev R. BERT Explained: State of the art language model for NLP // Towards Data Science. 2018. Vol. 10.
12. Yarushkina N.G., Moshkin V.S., Andreev I A. The sentiment-analysis algorithm of social networks text resources based on ontology // Proceedings of the ITNT 2020. 2020. Pp. 226–232.
13. Learning word vectors for sentiment analysis / A. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts // Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011. С. 142–150.
14. Loukachevitch N. V., Chetviorkin I. I. Open evaluation of sentiment analysis systems based on the material of the Russian language // Scientific and Technical Information Processing. 2014. Vol. 41. Pp. 370–376.
15. Посевкин Р.В., Бессмертный И.А. Применение сентимент-анализа текстов для оценки общественного мнения // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. №. 1. С. 169–171.
16. ГОСТ 12.0.002-2014 ССБТ. Термины и определения
17. ТОИ Р-45-084-01 Типовая инструкция по охране труда при работе на персональном компьютере
18. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования. [Электронный ресурс] Режим доступа: http://docs.cntd.ru/document/ 1200005187 – свободный (дата обращения: 27.05.2023)

19. СП 52.13330.2016. Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*. – М.: ИПК Изд-во стандартов, 2017. – 122 с.
20. ГОСТ 12.1.038-82 Система стандартов безопасности труда (ССБТ). Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов
21. СП 12.13130.2009. Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.
22. СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» // Электронный фонд правовой и нормативно технической документации [Электронный ресурс]. 2021. – Режим доступа: https://docs.cntd.ru/document/573500115 (дата обращения: 12.03.2023);