

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

Школа ИЯТШ

Направление подготовки 01.04.02 прикладная математика и информатика

Отделение школы (НОЦ) Экспериментальной физики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРАНТА

Тема работы
Алгоритм и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения

УДК 004.415.2.021:004.738:316.772.4:331.548

Обучающийся

Группа	ФИО	Подпись	Дата
0BM12	Гузеев Евгений Викторович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Семенов М.Е	к.ф.м.н		

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

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Горюнова Н.Н	к.э.н		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Сечин А.А	к.т.н		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП/ОПОП, должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Мерзликин Б.С	к.ф.м.н		

Планируемые результаты обучения по ООП

Код компетенции	Наименование компетенции
Профессиональные компетенции	
ПК(У)-1	Способен проводить научные исследования и получать новые научные и прикладные результаты самостоятельно и в составе научного коллектива
ПК(У)-2	Способен проводить поиск и анализ научной и научно-технической литературы по тематике проводимых исследований
ПК(У)-3	Способен разрабатывать и анализировать показатели качества информационных систем, используемых в производственной деятельности
ПК(У)-4	Способен планировать научно-исследовательскую деятельность, анализировать риски, управлять проектами, управлять командой проекта
ПК(У)-5	Способен к преподаванию математических дисциплин и информатики в образовательных организациях высшего образования
ПК(У)-6	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
Общепрофессиональные компетенции	
ОПК(У)-1	Способен решать актуальные задачи фундаментальной и прикладной математики
ОПК(У)-2	Способен совершенствовать и реализовывать новые математические методы решения прикладных задач
ОПК(У)-3	Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности
ОПК(У)-4	Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности с учетом требований информационной безопасности
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном языке, для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Инженерная школа ядерных технологий
Направление подготовки 01.04.02 «Прикладная математика и информатика»
Отделение экспериментальной физики

УТВЕРЖДАЮ:

Руководитель ООП

_____ Мерзликин Б.С.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

Студенту:

Группа	ФИО
0BM12	Гузееву Евгению Викторовичу

Тема работы:

Алгоритм и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения	
Утверждена приказом директора (дата, номер)	№34-108/с от 03.02.2023

Срок сдачи студентом выполненной работы:	6.06.2023
--	-----------

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Собрать данные в социальной сети вконтакте, в качестве признакового пространства использовать группы в которых состоит пользователь. Эти данные использовать для разработки алгоритма по выбору направления обучения.
---------------------------------	---

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Провести анализ литературы по теме диссертации. 2. Провести теоретическую основу и на основе нее построить алгоритм 3. Концепция старта-проекта 4. Социальная ответственность
Перечень графического материала	
Консультанты по разделам выпускной квалификационной работы (если необходимо, с указанием разделов)	
Раздел	Консультант
«Концепция стартап-проекта»)	Горюнова Н.Н
Социальная ответственность	Сечин А.А
Иностранный язык	Смирнова У.А

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		20.12.2022 г.

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0BM12	Гузеев Евгений Викторович		20.12.2022 г.

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Инженерная школа ядерных технологий

Направление подготовки 01.04.02 «Прикладная математика и информатика»

Отделение экспериментальной физики

Период выполнения весенний семестр 2022/2023 учебного года

Форма представления работы:

Магистерская диссертация

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	6.06.2023
--	-----------

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
05.02.2023	Выдача задания	5
14.02.2023	Обсуждение структуры работы	5
28.02.2023	Обзор литературы	10
16.03.2023	Сбор данных через API вконтакте	15
28.03.2023	Написание функции чистки текстовых данных	30
14.04.2023	Создание моделей классификации	15
14.05.2023	Сбора данных и создание моделей для других специальностей	10
31.05.2023	Написание пояснительной записки ВКР	10

СОСТАВИЛ:

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Семенов М.Е	к.ф.-м.н.		

СОГЛАСОВАНО:

Руководитель ООП

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Мерзликин Б.С.	к.ф.-м.н., доцент		

РЕФЕРАТ

Выпускная квалификационная работа _____ 126 _____ с., _____ 25 _____ рис., 18 _____ табл., 45 источников, _____ прил.

Ключевые слова: социальные сети, анализ данных, модель, классификация, признаковое пространство, сбор данных, алгоритм, машинное обучение.

Объектом исследования является анализ исследования принятия решений по выбору направления обучения.

Цель работы – _____ разработка алгоритма и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения, основанного на методах анализа социальных сетей и машинного обучения.

В процессе исследования был проведен анализ источников литературы по теме, разработаны структуры двух алгоритмов. Проведено сравнение точности этих алгоритмов.

В результате исследования было выяснено, что признаковое пространство пользователей вконтакте в виде списка групп достаточно хорошо подходит для выявления интересов пользователей.

Степень внедрения: реальна. На базе алгоритма планируется создать платформу по рекомендации направления образования.

Оглавление

ВВЕДЕНИЕ	8
1 ОБЗОР НАУЧНО-ТЕХНИЧЕСКОЙ ЛИТЕРАТУРЫ ПО ИССЛЕДОВАНИЮ	10
1.1 Цифровые следы в образовательной деятельности.....	10
2 ТЕОРЕТИЧЕСКАЯ ОСНОВА ИССЛЕДОВАНИЯ И ПОСТАНОВКА ЗАДАЧИ	16
2.1.1 Выбор метрики качества классификации	19
2.1.2 Подготовка текстовых данных для модели классификации	20
Векторное представление слов	20
2.2 Теоретическая база для алгоритма 2	37
2.2.1 Алгоритм градиентного бустинга CatBoost	38
3 КОНЦЕПЦИЯ СТАРТАП-ПРОЕКТА	43
4 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ	87
4.1 Правовые и организационные вопросы обеспечения безопасности	87
4.2. Производственная безопасность.....	90
4.3. Экологическая безопасность.....	95
4.4 Безопасность в чрезвычайных ситуациях	96
ЗАКЛЮЧЕНИЕ	101
СПИСОК ПУБЛИКАЦИЙ.....	106
ПРИЛОЖЕНИЕ А	107
ПРИЛОЖЕНИЕ Б ОСНОВНЫЕ ССЫЛКИ ПО РЕШЕНИЮ	124
ПРИЛОЖЕНИЕ В ОСНОВНЫЕ ФУНКЦИИ.....	125

ВВЕДЕНИЕ

Социальные сети в наше время играют важную роль в жизни людей. Люди узнают из социальных сетей новости, им показывают различную рекламу, многие формируют клубы по интересам, дате рождения, возрасту очень много взаимодействий между вытекает из полученной информации. В социальных сетях на основе собираемой информации основываются рекомендательные системы, которые рекомендуют пользователям определенный контент или рекомендуют пользователю определенного человека.

Социальные сети в современной жизни используются повсеместно. Бизнес, например, из социальной сети ищет свою целевую аудиторию, таргетируясь на которую, можно оптимизировать затраты на рекламный бюджет. Некоторые СМИ молниеносно могут распространить определенную новость, используя при этом вирусную составляющую новостного контента. В социальной сети можно выявлять деструктивные сообщества или сообщества, которые представляют угрозу в современном мире.

Так как люди почти ежедневно используют социальные сети и оставляют определенные социальные следы: интересы, репосты, музыкальные предпочтения, подписки. На основе этих данных можно получить примерный цифровой портрет человека.

Такой портрет можно использовать достаточно обширно от маркетингового взаимодействия до выявления социального поведения пользователей.

Сейчас в образовательной политике РФ идет тенденция на выявление более талантливых учащихся их профессиональную ориентацию и прогнозирование их профессиональных успехов. Традиционными признаками данных для решения этих задач используются данные, генерируемые пользователями в системах электронного обучения, а также психологические и социологические опросы. Однако, этих данных для выявления более полной картины мира учащегося и составления образовательных прогнозных моделей

недостаточно. Поэтому тут на помощь могут прийти социальные сети. Пользователь социальной сети, если ведет активную жизнь (круг общения, круг интересов, музыкальные предпочтения), то можно сказать, что он в достаточной степени показывает свои предпочтения и интересы. И вот эти предпочтения и интересы можно использовать как признаковое пространство для составления прогнозных моделей.

Такие прогнозные модели на основе интересов пользователя в социальной сети могут помочь с выбором специальности или направления обучения. При этом если человек учится и работает в соответствии со своими интересами, то образование для него пройдет легче и гораздо эффективнее.

Целью настоящей магистерской диссертации является разработка алгоритма и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения, основанного на методах анализа социальных сетей и машинного обучения.

Научная и практическая новизна и значимость работы:

В предложенном методе решена проблема выбора направления обучения абитуриентами, студентами и специалистами. На основе разработанного алгоритма будет построена цифровая платформа, которая будет рекомендовать пользователям наиболее релевантный вариант обучения в зависимости интересов пользователя социальной сети ВКОНТАКТЕ.

Для достижения этой цели необходимо проработать следующие задачи:

1. Привести обзор литературы по анализу социальных сетей и использованию данных как маркеров для выявления профессиональных интересов.
2. Сделать описание алгоритма и его описание.
3. Привести математический аппарат модели классификации данных
4. Реализовать данный алгоритм на нескольких специализациях
5. Сделать основные выводы и описать результаты исследования.

1 ОБЗОР НАУЧНО-ТЕХНИЧЕСКОЙ ЛИТЕРАТУРЫ ПО ИССЛЕДОВАНИЮ

1.1 Цифровые следы в образовательной деятельности

Изучение цифрового следа учащегося может улучшить аналитику и прогнозирование обучения несколькими способами:

Во-первых, цифровой след предлагает множество открытых пользовательских данных о личностных характеристиках, таких как когнитивные, мотивационные и психологические черты. Эти данные генерируются учащимся естественным образом, а не искусственно, например, с помощью тестов или опросов.

Во-вторых, анализ цифровых следов студента может помочь оценить его образовательный потенциал на начальном этапе обучения в университете. В отличие от существующих моделей, которые полагаются на текущую академическую успеваемость, анализ цифровых следов может предсказать успех студентов, которые только поступили в университет и еще не сдали свою первую сессию.

В-третьих, цифровые следы предоставляют новый способ оценки и анализа образовательных достижений, которые формально не фиксируются в рамках учебной программы университета. В результате растет потребность в аналитических инструментах и моделях, способных эффективно работать с цифровыми следами для повышения качества управленческих решений в образовательном процессе и образовательных траекторий студентов.

В 2019 году авторы Томского Государственного Университета [1] проводили исследование по созданию алгоритма поиска подходящих старшеклассников во ВКонтакте. По результатам работы алгоритм авторов определяет уровень креативности, интеллекта и мотивации.

Пример алгоритма представлен на рисунке 1.



Рисунок 1 – Пример алгоритма по поиску старшеклассников для университета [1]

Авторы [1] исследовали как текстовый контент на стенах пользователя, так и их подписки. В общем случае они исследовали порядка 9 тысяч образовательных сообществ из 1 млн развлекательных. Результаты исследования представлены на рисунке 2.

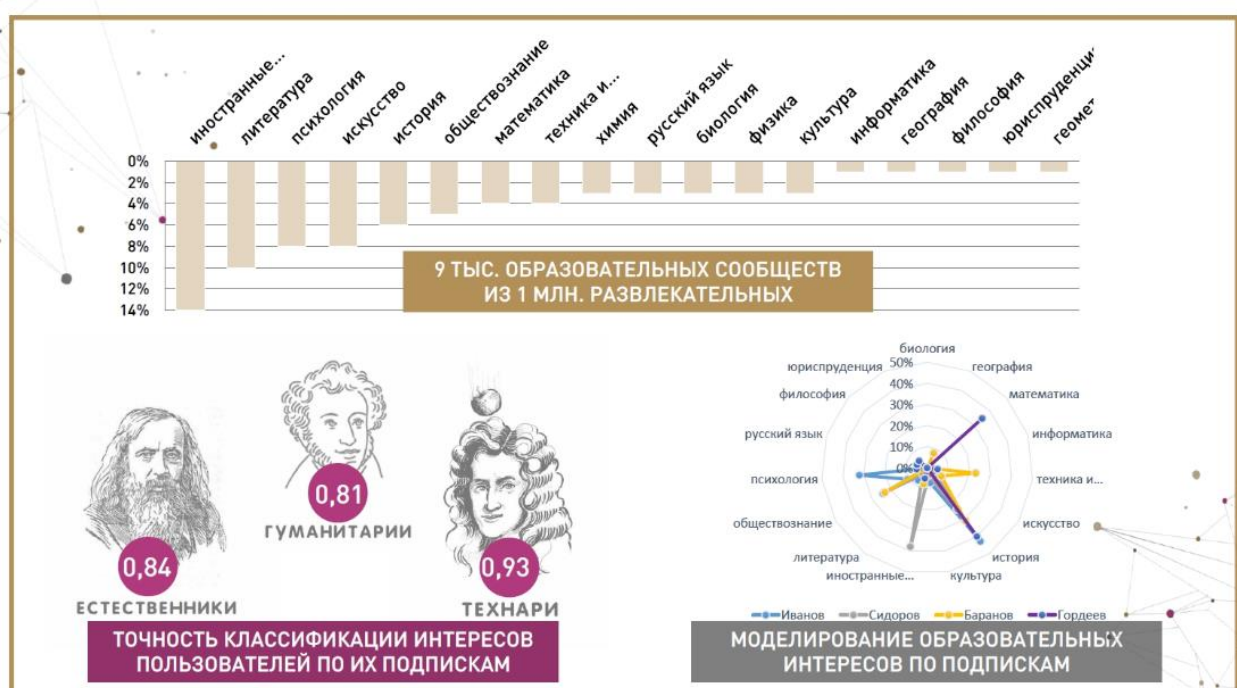


Рисунок 2 – Результаты анализа подписок пользователей [1]

Также они использовали в своей работе сетевой анализ для выявления уровней отличия креативности. Результаты представлены на рисунке 3.



Рисунок 3 – Результаты сетевого анализа [1]

Авторами на основе сетевого анализа был сделан вывод, что сети старшеклассников с низким уровнем креативности обычно состоят из 120–220 друзей, эти сети исключительно сегментированы, каждый граф состоит в среднем из 13–25 кластеров. Из них можно выделить 5–6 крупных сообществ, которые практически не связаны между собой, об этом также свидетельствует высокий коэффициент модулярности при низкой плотности сети [1].

Школьники с высоким уровнем креативности имеют склонность к формированию большого количества контактов, их сеть состоит в среднем из 260–400 друзей, эти сети имеют выраженный сегментарный характер. Однако различные кластеры имеют значительное количество связей между собой.

В источнике [2] авторами проводится исследование профессиональных интересов пользователя социальной сети на основе его данных, дополненных профессиональными опросами по выявлению интересов. По результатам исследования авторами была составлена прогнозная модель. Метрика ROC-AUC этой модели по разным направлениям изменялась от 0.7 для сферы управления до 0.85 до производственной сферы. Также авторы [2] показывают, что искусственный интеллект не может помочь с построением карьеры, но может помочь в осознании профессиональных интересов.

Примерно с 2015 года, начинается рост публикаций по теме анализа цифровых следов в образовательной деятельности. Например, в источнике [3]

автор показывает взаимосвязь между присутствием учащегося в социальной сети «ВКОНТАКТЕ» и его средним балом.

В статье [4] авторы рассматривают методологию выявления перспективных абитуриентов в социальной сети. Согласно исследованию, авторы выделили три направления для поиска перспективных абитуриентов по следующим направлениям: гуманитарные науки, точные науки, естественные науки. Все исследования авторы поделили на четыре направления [4]:

В первом направлении авторами была выдвинута гипотеза о взаимосвязи интересов школьников к контенту в социальной сети «ВКОНТАКТЕ». Во втором направлении авторы проверяли гипотезу о том, что персональные страницы пользователей могут отражать интересы абитуриента. В третьем направлении изучается потенциал социальных сетей для выявления важных психологических качеств абитуриента. В рамках четвертого направления предполагалось наличие особенностей в структуре социальных связей у одаренных старшеклассников [4].

Результаты их исследования показывают, методика предсказания образовательных интересов и признаков одаренности по подпискам пользователей дала наилучшую результативность. Также из выводов можно сделать, что одаренность абитуриентов в социальной сети достаточно трудноуловима из-за неравномерности контекстов подписок пользователя.

В источнике [5] авторы совместили использование цифровых следов, обучающихся с их следами в обучающейся среде LMS Moodle. Данные полученные из социальной сети и образовательной среды использовались авторами в качестве признакового пространства модели прогнозирования образовательных достижений студентов. Для построения модели ими были использованы современные библиотеки на основе алгоритма градиентного бустинга следующих библиотек: CatBoost и LightGBM. Также стоит отметить подход авторов, в котором они считали высокими образовательными интересами. Авторами было предложено считать, что высокими

образовательными достижениями обладают студенты, чей средний балл выше границы 65-го перцентиля, а низкими – ниже границы 35-го перцентиля.

В источнике [6] рассматривается предсказание персонального успеха пользователя на основе активности в социальной сети. Авторами пользователи с помощью алгоритма ближайших соседей были разбиты на 5 кластеров в зависимости от успешности. В ходе работы они дополняли входные данные включая, количество друзей, количество подписчиков, количество записей на стене.

В ходе вычислительных экспериментов [6] было обнаружено было обнаружено положительное влияние на точность нейронной сети при добавлении большего количества входных параметров для обучения нейронной сети. Кроме того, было подтверждено, что распределение исходной выборки людей на кластеры успеха является правильным, так как наблюдается монотонное увеличение различий с увеличением диапазона между двумя выделенными кластерами. Разделение кластера 5 (наиболее профессионально успешные люди) от всех остальных не показало хороших результатов.

Вектор характеристик пользователя

Для каждого пользователя сети будет собственный вектор характеристик исходя из его интересов. Вектор характеристик — это вектор заданной размерности, который будет характеризовать пользователя социальной сети. Для его построения используются, весовые коэффициенты, отвечающие за вклад каждого элемента из имеющихся наборов информации в конечный результат. Итоговые значения этих величин, полученные при учете всей доступной информации о пользователе суммированием с нормализацией, введенной для учета таких аспектов как возможность предоставления различного количества информации пользователями, разный уровень значимости одних результатов относительно других и прочих, образуют необходимый для сравнения вектор. Каждый элемент вектора соответствует некоторым интересам пользователя.

Опять же не всегда интересы пользователя совпадают с его выбором. Тут можно выявить два фактора выбора– осознанный фактор и неосознанный фактор. Неосознанный фактор – это фактор выбора, в которых человек при выборе университета не задействует собственные интересы.



Рисунок 1 – Факторы выбора

Чтобы исключить влияние неосознанных факторов необходимо провести кластеризацию интересов пользователей и минимизировать влияние интересов не на что не влияющих. К таким интересам можно отнести такие темы, как: юмор, игры, аукционы, объявления, городские новости и.т.д. Так как в социальной сети «ВКОНТАКТЕ» нет четкой кластеризации групп, сообществ, а есть только ее описание. В этом случае ставится задача кластеризация групп по тематикам на основании извлечения смысла из ее описания, то есть тематическое моделирование. Такой подход достаточно трудозатратный по времени и его результат нужно проверять.

Стоит заметить, что признаковое пространство контента пользователей представляет собой текстовые признаки. Поэтому целесообразно в разрабатываемом алгоритме рассматривать модели классификации на основе текстовых данных.

2 ТЕОРЕТИЧЕСКАЯ ОСНОВА ИССЛЕДОВАНИЯ И ПОСТАНОВКА ЗАДАЧИ

Контент в социальных сетях состоит из видео, картинок, музыки и текстового описания. При чем текстовое описание присутствует везде: в описаниях, в комментариях, в названиях и.т.д. Современные методы машинного обучения не умеют работать с текстом. Поэтому целесообразно репрезентовать текстовые данные так, чтобы можно было использовать методы машинного обучения.

Ключевой гипотезой в разработке алгоритма является то, что человек выбирает свое образование согласно своим интересам. Для проверки этой гипотезы были собраны данные по API ВКОНТАКТЕ 786 человек, обучающихся по направлению «Прикладная математика и информатика» в России за период с 2013 года. Такой ранний период был взят для расширения объема выборки. Результаты сбора данных показаны на рисунке 1.

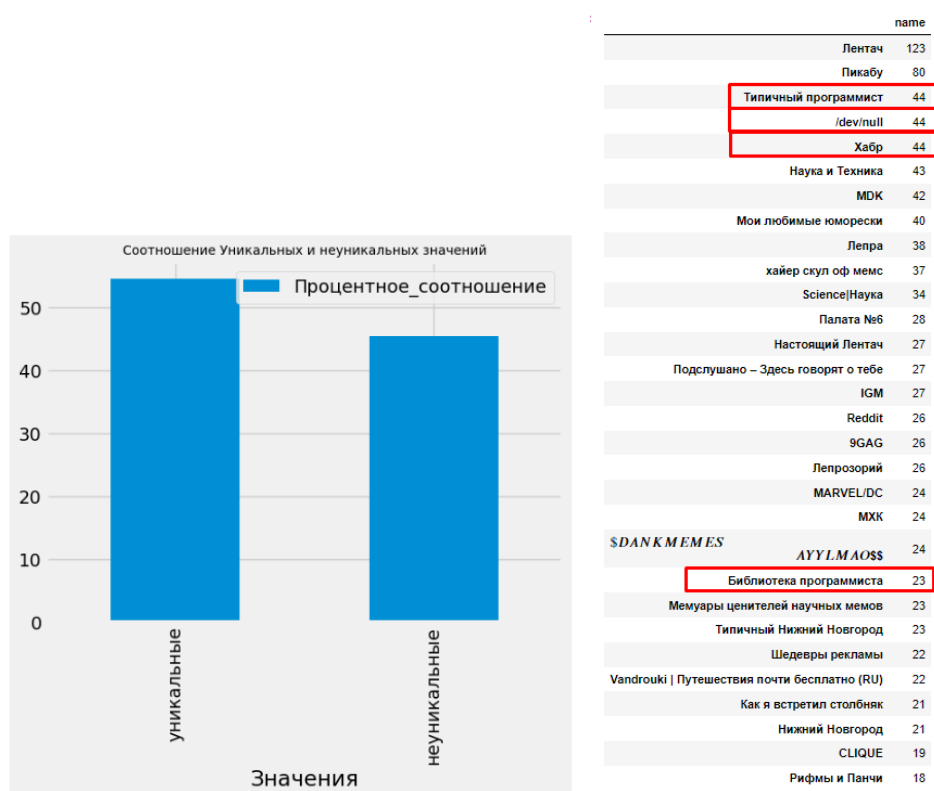


Рисунок 1 – Результат сбора данных

Как видно из рисунка, что 45,5% от всей выборки состоят в одинаковых группах. Также среди интересов в топе 5 групп, находятся группы, имеющие непосредственное отношение к программированию (помечены красным цветом).

Отсюда вытекают две ключевые идеи для создания прогнозных алгоритма:

1. Ручная разметка данных. Создание и обучение модели на основе этих данных.

В таком подходе есть как свои достоинства, так и свои недостатки. К достоинствам можно отнести более точную разметку данных, согласно экспертной оценке. Данные размечались на 4 класса 0 – группа не имеет отношения, а 3 – группа имеет прямое отношение к интересующей целевой аудитории. Распределение по классам (в долях): 0 – 0,52; 1 – 0,37; 2 – 0,073; 3 – 0,037.

К недостаткам относится большие затраты по времени для создания одной модели. Эти затраты можно сократить, через тематическое моделирование тем, но необходимо смотреть на метрики классификации. Также большая сложность

Структурная схема такого алгоритма представлена на рисунке 2.

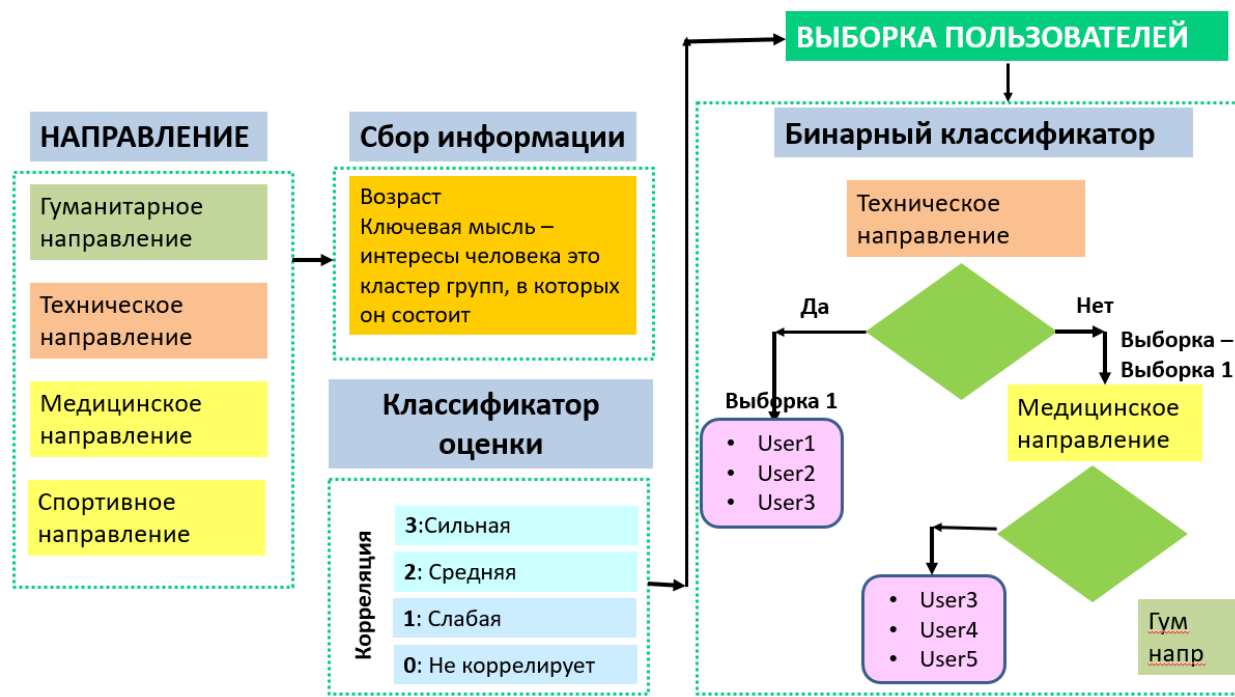


Рисунок 2 – Структурная схема алгоритма 1

Алгоритм работает следующим образом:

1. На вход алгоритма подается выборка пользователей. Этой выборке проставляются баллы. Баллы проставляются путем предсказания мультиклассификационной модели к отношению направления. 1 слабо коррелирует, 4 сильно коррелирует с направлением.

2. После этого выборка пользователей подается в стек с предообученными моделями. И выборка пользователей распределяется по направлениям.

2.1 Теоретическая база для алгоритма 1.

Рассмотрим алгоритмы классификации из простого конвейера классификации. Простой конвейер представлен на рисунке 3.

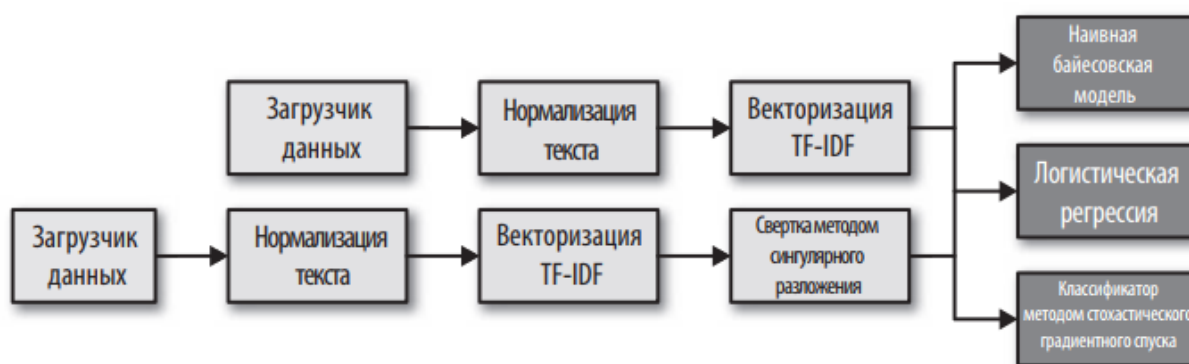


Рисунок 3 – Простой конвейер классификации

2.1.1 Выбор метрики качества классификации

В случае классификации самыми распространенными метриками являются:

- точность–полнота;
- верность;
- верность с учетом стоимости;
- площадь ROC кривой AUC.

Точностью (precision) называется отношение истинно положительных предсказаний к общему числу положительных предсказаний:

$$precision = \frac{TP}{TP + FP}$$

Полнотой (recall) называется отношение истинно положительных предсказаний к общему числу положительных примеров:

$$recall = \frac{TP}{TP + FN}$$

Верность (accuracy) – это число правильно классифицированных примеров, поделенное на общее число классифицированных примеров. В терминах матрицы неточностей верность равна

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

F-мера

Хотя точность и полнота являются очень важными метриками, сами по себе они не дадут вам полной картины, особенно в условиях дисбаланса классов. Одним из способов подытожить их является F-мера (F -measure), которая представляет собой гармоническое среднее точности и полноты, то есть обратное значение от среднего значения обратных значений точности и полноты:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

2.1.2 Подготовка текстовых данных для модели классификации

Векторное представление слов

В анализе данных текст называется корпусом, а каждое наблюдение в корпусе текста называется документом. Одно из самых простых вариантов представления текстовых данных является «мешок слов». В этом представлении и корпуса удаляется вся структура текста и используется только частота слова в тексте. Представление «мешок слов» представлено на рисунке 4.

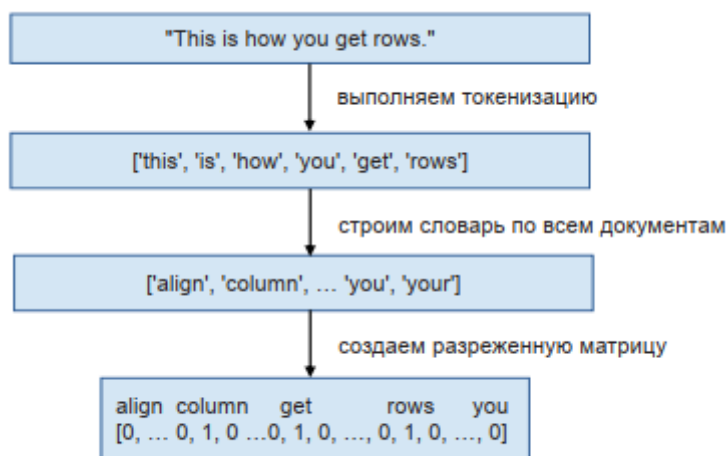


Рис. 58 Представление «мешок слов»

Рисунок 4 – Представление bag of words [7]

Токенизация текста включает в себя следующие этапы:

- Удаление стоп-слов. Из текста удаляются слова, которые не несут смысловой нагрузки. Такими словами могут предлоги, местоимения, частицы, знаки пунктуации, различные смайлы.
- Привести весь текст к одному регистру.
- После этого подготовленный текст разбивается на слова.

После токенизации необходимо провести лемматизацию слов, то есть привести их к нормальной словоформе. Пример лемматизации слов представлен на рисунке 5.



Рисунок 5 – Лемматизация слов

В некоторых ситуациях используется стемминг. Это извлечение основы слова. Например лесной → лес.

TF-IDF пространство

В алгоритме TF-IDF слову присваивается вес в зависимости от того, насколько часто оно встречается в наборе документов. Если в одной категории много документов с одним и тем же признаковым словом, значит, это признаковое слово имеет хороший потенциал для ее представления. Признаковому слову "сильный" следует придать больший вес. Однако если признаковое слово является общим для всех категорий, независимо от того, сколько раз оно встречается в каждом наборе документов, ему следует придать меньший вес.

Для того чтобы получить вес слова, частота слова умножается на обратную частоту документа. Это частота встречаемости слова в документе внутри класса, а обратная частота документа - это информация о распределении слова, встречающегося во многих категориях. В формуле

алгоритма TF-IDF есть две части: TF и IDF. Частота слов в классе представлена термином "частота слов", сокращенно "TF". Расчет TF-IDF представлен в следующем уравнении:

$$tf-idf(t, d, D) = tf(t, d) \cdot idf(t, D) = \frac{n_t}{\sum_k n_k} \cdot \log \frac{|D|}{|\{D_i \in D \mid t \in D_i\}|},$$

где, n_t — количество вхождений слова t в тексте, а n_k — общее количество слов в этом тексте(документе) из коллекции D , в которых встречается слово t .

Кодирование слов

Если в каком-либо признаковом пространстве в категориальных признаках имеются текстовые данные, то целесообразно использовать следующие подходы к кодированию:

1. Найти и заменить. Такой подход используется, когда количество значений относительно немного. Например, «четыре» меняются на 4.
2. Label encoding (кодирование метки). Это простое преобразование признака в число. Например, в данных у нас 125 уникальных категориальных значений, значит в результате кодирования будет 125 меток от 1 до 125. Пример кодировки представлен на рисунке 6.
3. Унитарное кодирование. Такая стратегия состоит в том, чтобы преобразовать значения каждой категории в новый столбец. Пример кодировки представлен на рисунке 7.

Переменная Class		Переменная Class	Кодировка LabelEncoder для переменной Class
A	0	B	1
B	1	A	0
C	2	C	2
D	3	A	0
		A	0
		D	3
		D	3
		B	1

Рисунок 6 – Пример Label Encoding [8]

State	Washington	Arizona	Nevada	California	Oregon
Washington	1	0	0	0	0
Arizona	0	1	0	0	0
Nevada	0	0	1	0	0
California	0	0	0	1	0
Oregon	0	0	0	0	1

Рисунок 7 – Пример унитарного кодирования [9]

CatBoost (целевое кодирование)

Если у признака имеется большое количество категорий, например id пользователя, то унитарное кодирование приведет к большому количеству новых признаков. Чтобы решить такую проблему, можно кластеризовать или объединить эти категории в ограниченное количество. Одним из путей решения данной проблемы является представление категориальных переменных с помощью статистик зависимой переменной. Например, мы можем представить каждую категорию в виде среднего значения зависимой переменной в данной категории [9].

Чтобы заменить категорию x_k^i -го обучающего наблюдения одним количественным признаком, равным некоторой статистике зависимой переменной $\hat{x}_k^i$. Обычно эта статистика оценивает ожидаемое значение

зависимой переменной y , обусловленное конкретной категорией:

$$\hat{x}_k^i \approx E(y | x^i = x_k^i) \text{ [8].}$$

Самый простой подход - оценить $E(y | x^i = x_k^i)$ как среднее значение зависимой переменной y по всем обучающим наблюдениям, принадлежащим к одной и той же категории. Эта оценка будет шумной для редких категорий и поэтому ее обычно сглаживают некоторым априорным значением P :

$$\hat{x}_k^i = \frac{\sum_{j=1 \{ x_j^i = x_k^i \}}^n y_j + a \cdot P}{\sum_{j=1 \{ x_j^i = x_k^i \}}^n 1 + a},$$

где $a > 0$ – параметр. В качестве P обычно берут среднее значение зависимой переменной в наборе данных. Проблема такого жадного подхода - утечка информации о зависимой переменной, признак $\hat{x}_k^i$ вычисляется с помощью y_k , значения зависимой переменной для x_k . Это приведет к условному смещению распределение $\hat{x}^i | y$ будет разным для наблюдений обучающего и тестового наборов [8]. Следующий пример проиллюстрирует, насколько данное обстоятельство может повлиять на обобщающую способность обученной модели. Предположим, что i -тый признак является категориальным, все его значения уникальны, и, решая задачу классификации, для каждой категории A мы получаем $P(y = 1 | x^i = A) = 0,5$. В обучающем наборе мы получаем $\hat{x}_k^i = \frac{y_k + aP}{1 + a}$, таким образом, достаточно выполнить одно расщепление с помощью порога $t = \frac{0,5 + aP}{1 + a}$, чтобы идеально классифицировать все обучающие наблюдения [9]. Однако полученная модель для всех тестовых наблюдений предсказывает 0, если $P < t$, и предсказывает 1 в противном случае, таким образом, в обоих случаях получаем правильность

0,5. Поэтому сформулируем следующее желаемое условие для вычисления статистики зависимой переменной:

Условие 1 $E(\hat{x}^i | y = v) = E(\hat{x}_k^i | y_k = v)$, где (x_k, y_k) – это k -ое обучающее наблюдение. В примере, приведенном выше, $E(\hat{x}_k^i | y_k) = \frac{y_k + aP}{1 + a}$ и $E(\hat{x}^i | y) = P$ различны. Существуют несколько способов избежать этого условного смещения. Главная идея, лежащая в их основе, заключается в том, чтобы вычислить статистику зависимой переменной для x_k на подмножестве примеров $D_k \subset D \setminus \{x_k\}$, исключая x_k :

$$\hat{x}_k^i = \frac{\sum_{x_j \in D_k \{x_j^i = x_k^i\}} y_j + a \cdot P}{\sum_{x_j \in D_k \{x_j^i = x_k^i\}} 1 + a}, \quad (1)$$

В CatBoost при вычислении статистики зависимой переменной для каждого наблюдения используется лишь имеющаяся история. На обучающем наборе выполняется случайная перестановка и теперь для интересующего наблюдения считается статистика зависимой переменной на основе всех наблюдений, которые в данной перестановке расположились выше наблюдения, которые встретились «раньше» «нашего» наблюдения и по отношению к нему являются «историей». Следует обратить внимание, если использовать только одну случайную перестановку, то у предшествующих обучающих наблюдений статистика зависимой переменной будет иметь гораздо большую дисперсию, чем у последующих обучающих наблюдений. Поэтому CatBoost выполняет несколько перестановок.

Рассмотрим, как происходит преобразование категориальных признаков в количественные при решении задачи классификации. Оно происходит в три этапа.

- 1) CatBoost принимает на вход набор значений признаков и значений зависимой переменной (т.е. моделируемых значений).

- 2) Наблюдения несколько раз случайным образом перемешиваются, и создается несколько случайных перестановок набора данных.
- 3) Все категориальные признаки преобразовываются в количественные по формуле:

$$avg_target = \frac{countInClass + prior}{totalCount + 1},$$

где

countInClass - сколько раз значение зависимой переменной равно 1 для наблюдений с данной категорией признака (учитываются только наблюдения с данной категорией признака до интересующего нас наблюдения, при этом расчеты выполняются в порядке следования наблюдений после перемешивания, то есть интересующее нас наблюдение в расчетах не участвует);

prior - исходное значение числителя, определенное начальными параметрами (например, 0,05);

totalCount - общее количество наблюдений до интересующего нас наблюдения, для которых категория признака совпадает с категорией признака для текущего наблюдения. Еще раз заметьте, интересующее нас наблюдение в расчетах не участвует. Вычисление значений будущей количественной переменной для категориального признака представлено на рисунке 8

я	Признак ₁	Признак ₂	Признак ₃	Зависимая переменная	Вычисление значений
1	4	53	Рок	0	$(0 + prior) / (0 + 1) = 0,05$
2	3	55	Инди	0	$(0 + prior) / (0 + 1) = 0,05 / 1 = 0,05$
3	2	40	Рок	1	$(0 + prior) / (1 + 1) = 0,05 / 2 = 0,025$
4	5	42	Рок	1	$(1 + prior) / (2 + 1) = 1,05 / 3 = 0,35$
5	5	34	Поп	1	$(0 + prior) / (0 + 1) = 0,05 / 1 = 0,05$
6	2	48	Инди	1	$(0 + prior) / (1 + 1) = 0,05 / 2 = 0,025$
7	2	45	Рок	0	$(2 + prior) / (3 + 1) = 2,05 / 4 = 0,5125$

Рисунок 8 – Вычисление значений будущей количественной переменной для категориального признака

2.1.3 Модели для классификации данных

Для задач мультиклассовой классификации достаточно хорошо подойдет **логистическая регрессия**, но при ее использовании необходимо заменить сигмоидную функцию функцией **softmax**:

$$P(y = k | X) = \frac{e^{\beta \cdot x_i}}{\sum_{j=1}^K e^{\beta \cdot x_j}},$$

где $P(y = k | X)$ – вероятность, что целевое значение i -го наблюдения y_i является классом k ; K – Любое количество классов

Алгоритм ближайших соседей можно распространить на задачи мультиклассовой классификации. Для этого необходимо отыскать k -ближайших соседей данных для входа X и вернуть класс, к которому принадлежит больше всего данных среди k [2].

Более формально вероятность x_i некоторого класса j имеет вид:

$$\frac{1}{k} \sum_{i \in v} I(y_i = j),$$

где v — k -е наблюдение в окрестности x_i ; y , — класс i -го наблюдения; I — индикаторная функция (т. е. 1 — истинно, 0 — в противном случае).

Наивный байесовский классификатор основан на теореме Байеса и предполагает, что каждый класс имеет свой набор параметров. В алгоритме вычисляется вероятность появления класса в выборке. Затем для нового объекта вычисляются вероятности его принадлежности к каждому классу на основе формулы Байеса. Алгоритм может использоваться в задачах мультиклассовой классификации если предположение о независимости признаков не нарушается.

В частности, наивный байесов классификатор основан на следующей формуле:

$$P(y | x_i, \dots, x_j) = \frac{P(y | x_i, \dots, x_j) \cdot P(y)}{P(x_i, \dots, x_j)},$$

$P(y|x_1, \dots, x_j)$ называется апостериорным значением и является вероятностью того, что наблюдение принадлежит классу y при условии, что это наблюдение имеет значения j признаков $x_1, \dots, x_j$;

$P(y|x_1, \dots, x_j)$ называется правдоподобием и является правдоподобной оценкой вероятности, что наблюдение имеет значения признаков $x_1, \dots, x_j$ при условии, что дан их класс y ;

$P(y)$ называется априорной вероятностью и является нашей субъективной оценкой вероятности класса y перед рассмотрением данных;

$P(x_1, \dots, x_j)$ называется предельной вероятностью.

Алгоритм случайного леса. Случайный лес состоит из нескольких деревьев решений, каждое из которых обучается на случайной подвыборке данных. Каждое дерево строится путем разбиения на две или более подгруппы на основе определенного признака. Когда новый объект поступает на вход, он проходит через каждое дерево, и каждое дерево выдает свой прогноз. В результате объекту присваивается класс, который получается путем голосования всех деревьев.

Итоговая формула классификатора будет выглядеть:

$$a(x) = \frac{1}{N} \sum_{i=1}^N b_i(x),$$

где

N – количество деревьев;

I – счетчик для деревьев;

b – решающее дерево;

x – выборка.

Главная идея **градиентного бустинга** заключается в том, что каждое следующее решающее дерево, учитывает ошибки предыдущих деревьев и пытается исправить их. Таким образом постепенно улучшается качество предсказаний.

Использование моделей-трансформеров для задач мультиклассовой классификации текста

Эффективность моделей-трансформеров в различных задачах NLP, включая генерацию и классификацию текстов, вызывает значительный интерес. Механизм внимания является важнейшим компонентом этих моделей. Однако еще до появления трансформеров механизм внимания использовался для улучшения обычных моделей глубокого обучения, таких как RNN. Чтобы получить полное представление о трансформерах и их важности в NLP, необходимо сначала изучить механизм внимания [12].

Этот механизм основан на том факте, что модели на основе RNN, такие как GRU или LSTM, имеют информационное узкое место для таких задач, как нейронный машинный перевод (neural machine translation, NMT). Эти модели на основе энкодера-декодера получают ввод в виде token-id и подвергают его рекуррентной обработке (энкодер). После этого обработанное промежуточное представление подается в другой рекуррентный блок (декодер) для извлечения результатов [12].

Чтобы устранить это узкое место и согласовать декодер с энкодером, был предложен механизм внимания, в котором используются веса, присвоенные промежуточным скрытым значениям. Эти веса определяют количество внимания, которое модель должна уделять входным данным на каждом этапе декодирования. Схема типичного механизма внимания представлена на рисунке 9 [11].

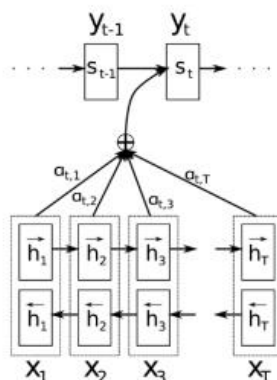


Рисунок 9 – Схема типичного механизма внимания [12]

Прежде чем переходить к механизмам внимания на основе взвешенных скалярных произведений, необходимо разобраться в том, что такое самовнимание (self-attention). Механизм самовнимания, схематически представленный на рисунке 10, является базовой формой механизма, взвешенного самовнимания (scaled self-attention) [12].

Этот механизм получает матрицу ввода X и производит оценку внимания ее элементов. В нашем случае X представляет собой матрицу 3×4 , где 3 – количество токенов, а 4 – размер представления. Матрица Q на рисунке 10 называется запросом (query), K – ключом (key), а V – значением (value). Эти матрицы получаются в результате перемножения X и матриц, помеченных символами θ , Φ и g соответственно. Перемножение матриц запроса (Q) и ключа (K) дает матрицу весов внимания. Этот механизм также можно рассматривать как базу данных, в которой мы используем запрос и ключи, чтобы получить численную оценку связей между различными элементами. Перемножение весов внимания и матрицы V дает конечный результат механизма внимания этого типа. Основная причина, по которой он называется самовниманием, – это его единственный вход X ; матрицы Q , K и V вычисляются из X .

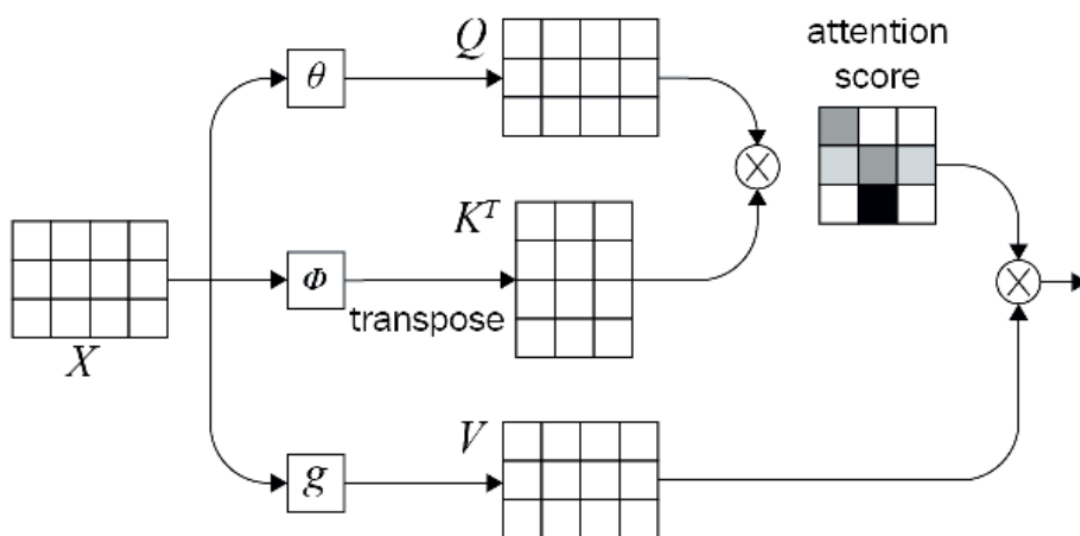


Рисунок 10 – Математическое представление механизма внимания [12]

Модели-трансформеры учитывают аннотации энкодера и скрытые значения из прошлых слоев. Архитектура трансформера не имеет рекуррентного пошагового потока; вместо этого в ней применяется позиционное кодирование, чтобы иметь информацию о позиции каждого токена во входной последовательности. На слои в первой части энкодера подаются входные данные, представляющие собой объединенные значения представлений (инициализированные случайным образом) и фиксированные значения позиционного кодирования; они распространяются по архитектуре, как показано на рисунке 11 [12].

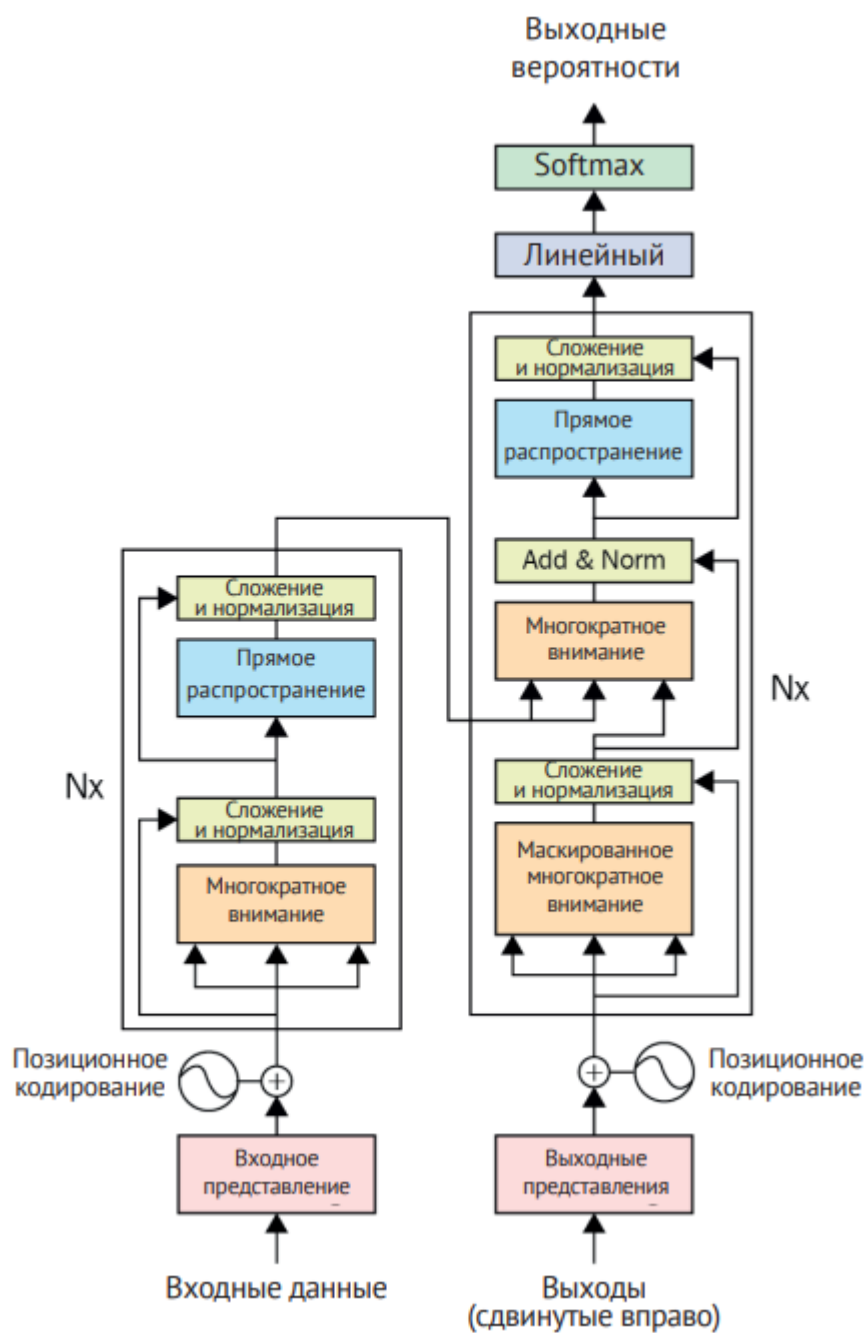


Рисунок 11 – Обобщенная модель трансформеров [12]

Тематическое моделирование

Существует несколько существующих алгоритмов, которые можно использовать для тематического моделирования. Наиболее распространенными из них являются латентно-семантический анализ (LSA/LSI), вероятностный латентно-семантический анализ (pLSA) и латентное распределение Дирихле (LDA). Для тематического моделирования в работе использовался алгоритм LDA.

LDA - это генеративная вероятностная модель, которая предполагает, что каждая тема является смесью над базовым набором слов, а каждый документ является смесью над набором вероятностей тем. Принцип алгоритма представлен на рисунке 12.

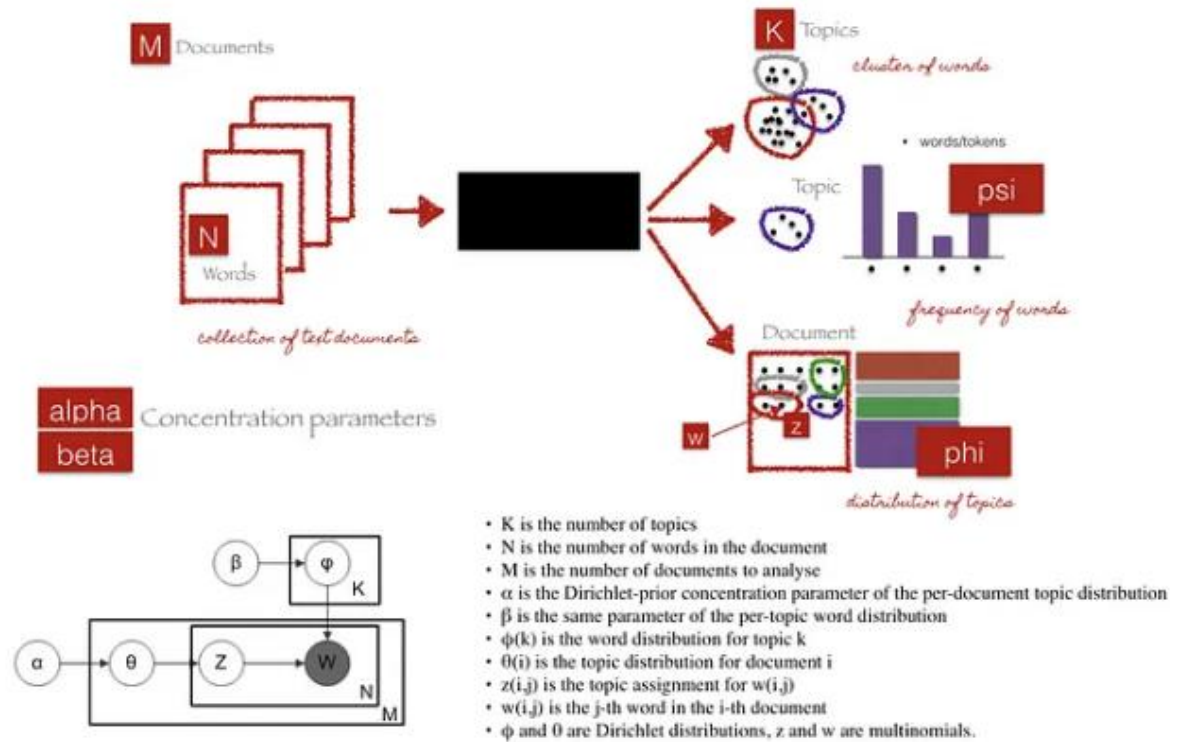


Рисунок 12 – Принцип алгоритма LDA [13]

Алгоритм работает следующим образом:

На первом шаге для каждого документа d выбирается случайный вектор распределения тем θ_d из распределения Дирихле с параметром α . На втором шаге выбирается тема t_{di} (в классической модели LDA количество тем фиксировано изначально) из мультиномиального распределения с параметром θ_d . Наконец согласно выбранной теме t_{di} выбирается слово w_{di} из распределения ϕ_t , которое является распределением Дирихле с параметром β . Таким образом, порождающая модель слова w из документа d представляется в виде:

$$p(w | d, \theta, \phi) = \sum_t p(w | t, \phi_t) p(t | d, \theta_d),$$

$\theta \sim \text{Dir}(\alpha)$, $\phi \sim \text{Dir}(\beta)$, где α и β — задаваемые так называемые гиперпараметры распределения Дирихле.

В ходе использования моделей мультиклассовой классификации текстов были получены следующие значения по метрике качества F_1 (таблица 1).

Таблица 1 – Сравнение метрики качества F_1 для задач мультиклассовой классификации текста

Алгоритм	Метрика F_1
Логистическая регрессия	0,729
Случайный лес	0,709
Наивный Байес	0,706
Ближайшие соседи	0,726
Градиентный бустинг (CatBOOST)	0,771
Трансформер с предобученной моделью:	
<i>rubert tiny 'cointegrated/rubert-tiny</i>	0,668
<i>distilbert-base-multilingual-cased</i>	0,691
<i>cointegrated/rubert-tiny2</i>	0,716

У трансформера с предобученной моделью лучше всех справилась модель *cointegrated/rubert-tiny2*. Тут можно достичь более лучших результатов, задействовав другие предобученные модели, но на некоторые модели требуется большой ресурс памяти.

Что касается моделей классификации лучше всех с задачей классификации текста справился алгоритм градиентного бустинга CatBoost. Также стоит отметить, что CatBoost удобен тем, что не надо преобразовывать текстовые данные, необходимо указать какие данные являются текстовыми, какие категориальными.

Приведем пример алгоритма для классификации двух специальностей прикладная математика и лечебная физкультура. Стоит отметить, что в качестве признакового пространства использовались 10 групп, в которых состоит пользователь. Для каждой группы были взяты данные по тематике, описанию и названию группы.

1. Загрузка сырых данных по специальности прикладная математика и лечебная физкультура

	id	description	activity	name
0	61674504	Официальная страница российской Суперлиги	Спортивная организация	Суперлига
1	61674504	РФС — общероссийская общественная организация,...	Спортивная организация	Российский футбольный союз РФС
2	61674504	Официальный аккаунт ЮФЛ-девушки и ЮФЛ-девочки.	Футбол	ЮФЛ-девушки ЮФЛ-девочки
3	61674504	Единственный турнир, где Калининград встречает...	Спортивная организация	ФНЛ Лучшая лига мира
4	61674504	Официальная страница женской сборной России по...	Футбольная команда	Женский футбол России

	id	description	activity	name
0	34295316	Присылайте свои анекдоты в "Предложите новость...	Юмор	Азохнвей! Бесплатные щикарные анекдоты
1	34295316	Познавательные мемы, широкий охват тем, уникал...	Юмор	Мемуары ценителей научных мемов
2	34295316	На Хабр Q&A можно получить ответ на вопрос по ...	Программирование	Хабр Q&A
3	34295316	Официальный канал сервиса РБК Инвестиции для т...	Финансы	РБК Инвестиции

Рисунок 13 – Необработанные данные для специальностей лечебная физкультура и прикладная информатика

2. Чистка данных и подготовка к обучению. При чистке данных используется удаление пунктуации, приведение к одному регистру, удаление гиперссылок, смайлов, различных символов. Размеченные данные представлены на рисунке 14.

	description_processed	activity_processed	name_processed	target
0	присылать анекдот предложить новость группа ре...	юмор	азохнвеать бесплатный щикарный анекдот	1
1	познавательный мем широкий охват уникальный ид...	юмор	мемуары ценитель научный мем	1
2	хабра q a ответ любой тема it участник сообщес...	программирование	хабра q a	4

Рисунок 14 – Размеченные данные по прикладной математике

3. Обучаем модель, предсказываем темы топиков по специальности лечебная физкультура.

	id	description_processed	activity_processed	name_processed	target
0	61674504	официальный страница российский суперлига	спортивный организация	суперлига	2
1	61674504	рфс общероссийский общественный организация за...	спортивный организация	российский футбольный союз рфс	2
2	61674504	официальный аккаунт юфл девушка юфл девочка	футбол	юфл девушка юфл девочка	2
3	61674504	единственный турнир калининград встречаться ха...	спортивный организация	фнл хороший лига	2
4	61674504	официальный страница женский сборная россия фу...	футбольный команда	женский футбол россия	2
5	61674504		футбол	женский футбол	2
6	61674504	приветствовать дорога группа начинающий опытни...	метать	путешествие металлист	2
7	61674504	подвох	публичный страница	клуб любитель математика	1
8	61674504		юмор	смеяться слз d	1
9	61674504	заходить развлечься посмеяться забавный нелепы...	юмор	прикол смеяк	1
10	184665973	добро пожаловать один самый больший сообщество...	юмор	чёткий прикол	1

Рисунок 15 – Предсказанные темы групп

4. Преобразовываем данные, зависимую переменную класс обучения отмечаем как (1- прикладная математика, 2 – лечебная физкультура). Перемешиваем данные и обучаем модель.

id	target_1	target_10	target_2	target_3	target_4	target_5	target_6	target_7	target_8	target_9	target
67177	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1
120886	2.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	1
213176	4.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1
625027	1.0	2.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	1.0	1
662983	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	1
...	...	...	...	...	...	...	...	...	...	...	...
778146263	2.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	0
785400187	1.0	1.0	2.0	3.0	1.0	2.0	2.0	2.0	2.0	1.0	0
797245427	1.0	1.0	2.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	0
800829227	2.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	0
802248215	2.0	2.0	2.0	2.0	3.0	2.0	2.0	2.0	2.0	1.0	0

Рисунок 16 – Данные для подачи в модель

Пример тестирования алгоритма при использовании двух специальностей лечебная физкультура и прикладная математика представлен на рисунке 17.

	precision	recall	f1-score	support
0	0.57	0.80	0.67	188
1	0.75	0.50	0.60	227
accuracy			0.64	415
macro avg	0.66	0.65	0.63	415
weighted avg	0.67	0.64	0.63	415

Рисунок 17 – Определение метрик классификации двух специальностей при использовании

Что касается метрики классификации этого алгоритма, то она достаточно низка.

2.2 Теоретическая база для алгоритма 2

Что если, нам не нужно преобразовывать в векторное пространство все признаки, если алгоритм градиентного бустинга CatBoost работает хорошо с текстовыми данными. Тогда алгоритм будет работать следующим образом:

1. Создается стек предобученных моделей
2. На вход алгоритму подается ID пользователя, который хочет узнать свою специальность
3. С этого ID собирается данные по его 10 группам. Данные проходят очистку и предобработку. Делается репрезентация признакового пространства для подачи в модель.
4. Эти данные последовательно проходят стек предобученных моделей
5. На выход выходит наиболее вероятная специальность или направление обучения пользователя.

Алгоритм представлен на рисунке 18.

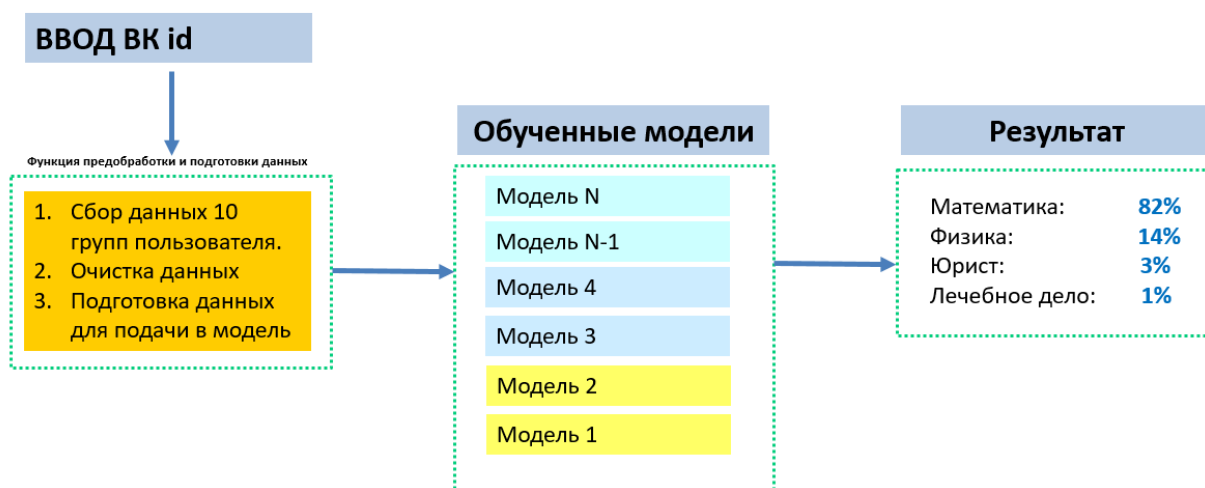


Рисунок 18 – Структурное решение для алгоритма 2

Преимущества данного алгоритма:

Скорость, сам алгоритм линейный и данные проходят через стек предобученных моделей.

Простота. В этом алгоритме мы не используем проставку баллов в зависимости от корреляции специальности, а подаем на вход сырые данные, которые проходят препроцессинг и подаются на вход в стек моделей.

Тогда рассмотрим подробнее алгоритм градиентного Бустинга CatBoost.

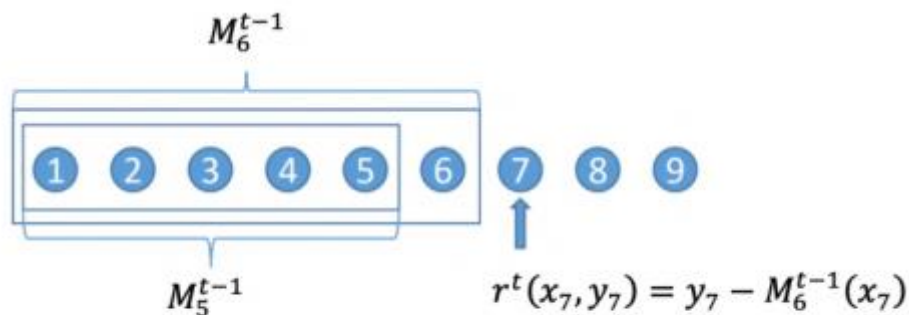
2.2.1 Алгоритм градиентного бустинга CatBoost

Двумя критически важными алгоритмическими достижениями, внедренными в CatBoost, являются реализация динамического бустинга, представляющая собой альтернативу классическому алгоритму бустинга и основанная на перестановках, а также инновационный алгоритм обработки категориальных признаков. Оба метода были созданы для борьбы со смещением прогноза, который вызван утечкой информации о зависимой переменной - проблемой, актуальной для всех существующих в настоящее время реализаций алгоритма градиентного бустинга [9].

CatBoost— это реализация градиентного бустинга, которая использует бинарные деревья решений в качестве базового алгоритма. Дерево решений представляет из себя модель, которая строится с помощью рекурсивного разбиения пространства признаков $\mathbb{R}^m$ на несколько непересекающихся областей (узлов дерева) в соответствии со значениями некоторых атрибутов расщепления a . Каждой итоговой области (листу дерева) присваивается значение, которое является оценкой зависимой переменной y в данной области, если решается задача регрессии, или спрогнозированной меткой класса, если задачи классификации. Таким образом, дерево решений h можно записать как

$$h(x) = \sum_{j=1}^J b_j \mathbb{1}_{\{x \in R_j\}},$$

где b_j - константа для R_j (среднее значение зависимой переменной или пропорции классов зависимой переменной в выборке наблюдений, относящейся к области R_j), R_j - это непересекающиеся области, соответствующие листьям дерева. CatBoost имеет два режима: динамический и обычный. Последний является стандартным алгоритмом градиентного бустинга над деревьями решений. Динамический бустинг представлен на рисунке 19 [8].



Принцип динамического бустинга

Алгоритм 1: Динамический бустинг

ввод: $\{(x_k, y_k)\}_{k=1}^n, I;$
 $\sigma \leftarrow$ случайная перестановка $[1, n];$
 $M_i \leftarrow 0$ для $i = 1 \dots n;$
для $t \leftarrow$ от 1 до I **выполнять**
 для $i \leftarrow$ от 1 до n **выполнять**
 $r_i \leftarrow y_i - M_{\sigma(t)-1}(x_i);$
 для $i \leftarrow$ от 1 до n **выполнять**
 $\Delta M \leftarrow \text{LearnModel}((x_j, r_j): \sigma(j) \leq i);$
 $M_i \leftarrow M_i + \Delta M;$
возвратить M_n

Рисунок 19–Динамический Бустинг [9]

Алгоритм 2: построение дерева в Catboost

ввод: $M, \{(x_i, y_i)\}_{i=1}^n, \alpha, L, \{\sigma_i\}_{i=1}^s$, Режим
 $\text{grad} \leftarrow \text{CalcGradient}(L, M, y);$
 $r \leftarrow \text{random}(1, s);$
 $G \leftarrow (\text{grad}_r(1), \dots, \text{grad}_r(n))$ для режима *Plain*;
 $G \leftarrow (\text{grad}_{r, \sigma_r(1)-1}(i) \text{ для } i = 1 \dots n)$ для режима *Ordered*;
 $T \leftarrow$ пустое дерево
для каждого этапа процедуры построения дерева «сверху-вниз» **выполнять**
 для каждого рассматриваемого разбиения c **выполнять**
 $T_c \leftarrow$ добавить разбиение c в T
 если режим == *Plain*, **то**
 $\Delta(i) \leftarrow \text{avg}(\text{grad}_r(p) \text{ для } p: \text{leaf}_r(p) = \text{leaf}_r(i))$ для всех $i;$
 если режим == *Ordered*, **то**
 $\Delta(i) \leftarrow \text{avg}(\text{grad}_{r, \sigma_r(i)-1}(p) \text{ для } p: \text{leaf}_r(p) = \text{leaf}_r(i), \sigma_r(p) < \sigma_r(i))$
 для всех $i;$
 $\text{loss}(T_c) \leftarrow \|\Delta - G\|_2$
 $T \leftarrow \text{argmin}_{T_c}(\text{loss}(T_c))$
если режим == *Plain*, **то**
 $M_{r'}(i) \leftarrow M_{r'}(i) - \alpha \text{avg}(\text{grad}_{r'}(p) \text{ для } p: \text{leaf}_{r'}(p) = \text{leaf}_{r'}(i))$ для всех $r', i;$
если режим == *Ordered*, **то**
 $M_{r', j}(i) \leftarrow M_{r', j}(i) - \alpha \text{avg}(\text{grad}_{r', j}(p) \text{ для } p: \text{leaf}_{r'}(p) = \text{leaf}_{r'}(i), \sigma_{r'}(p) \leq j)$
 для всех $r', i, j;$
возвратить T, M

Рисунок 20 – Построение дерева в Catboost [9]

Построение деревьев. CatBoost использует забывчивые деревья решений (их еще называют «небрежными»), в которых для всех узлов определенного уровня применяется одно и то же правило разбиения [9]. Такие деревья сбалансированы, в меньшей степени склонны к переобучению и позволяют

быстрее вычислить прогнозы. Процедура построения дерева в Catboost описана в алгоритме 2. Пример забывчивого дерева решений представлено на рисунке 21.

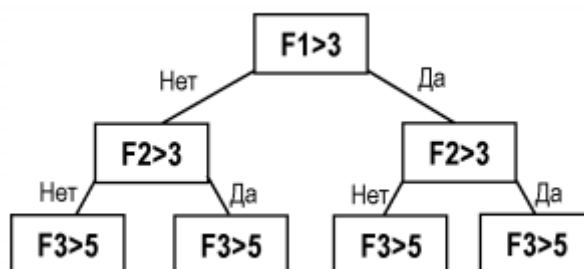


Рисунок 21– Пример забывчивого дерева решений

Посмотрим на результаты обучения по специальностям по алгоритму 2. Результаты представлены на рисунке 22.

	precision	recall	f1-score	support
0	0.82	0.90	0.86	177
1	0.92	0.85	0.89	239
accuracy			0.87	416
macro avg	0.87	0.88	0.87	416
weighted avg	0.88	0.87	0.87	416

Рисунок 22 – Результаты обучения в CatBoost на текстовых данных

В алгоритме 1 метрика F1 составляла 0.64, при этом метрика F1 по алгоритму 2 составила около 0.87. Прирост более 35%.

Тестирование алгоритма по другим специальностям представлено на рисунке 23.

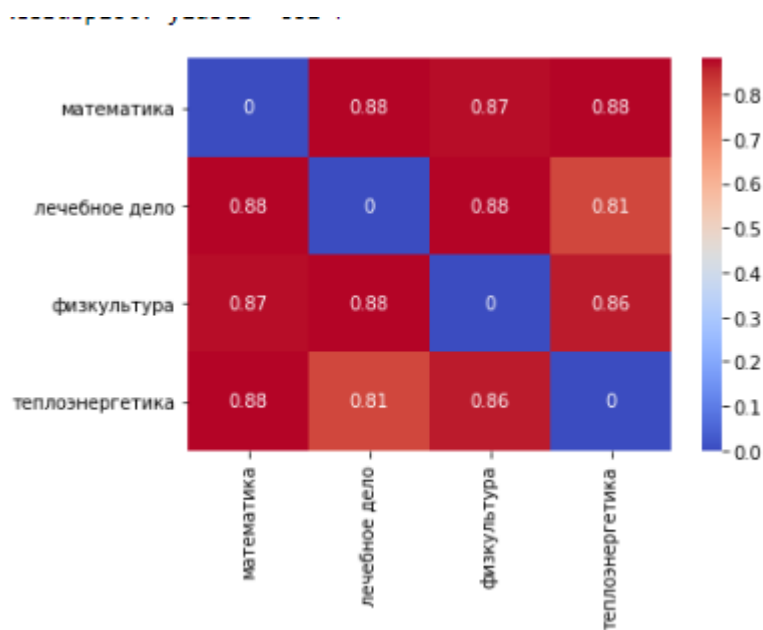


Рисунок 23 – Тестирование алгоритма по другим специальностям (метрика F1).

Как видно из рисунка 23 CatBoost достаточно хорошо, отличает специальности. Метрика F1 даже неполярных специальностей (теплоэнергетика и математика) составляет выше 0.8.

3 КОНЦЕПЦИЯ СТАРТАП-ПРОЕКТА

ПЛАТФОРМА «ПАРУС» – НАВИГАТОР ПО ВЫБОРУ ОБРАЗОВАНИЯ

3.1 Описание продукта как результата НИР

Выбор образования – это одно из самых важных решений, которое абитуриенты принимают в жизни. От этого выбора зависит не только их будущее, но и личная удовлетворенность и счастье. Однако, выбор образования может стать настоящей проблемой для многих людей.

При выборе образования необходимо учитывать множество факторов. Один из них – личные интересы и увлечения. Если студенты ВУЗов, онлайн-школ, техникумов будут заниматься тем, что им нравится, то учиться будет легче и результаты будут лучше. Поэтому перед выбором образования стоит задуматься, чем абитуриенты хотят заниматься в жизни и какие профессии их привлекают. Выбор образования может стать настоящей проблемой для многих людей.

Во-первых, существует множество различных видов образования, которые можно получить: от традиционного высшего образования до различных видов профессиональной подготовки и курсов. Каждый из них имеет свои преимущества и недостатки, и выбор может быть сложным.

Во-вторых, выбор образования может зависеть от многих факторов, таких как личные интересы, финансовые возможности, доступность образовательных программ и т.д. Некоторые люди могут ограничиваться в выборе из-за ограничений в финансах или доступности программ.

В-третьих, многие люди сталкиваются с давлением со стороны родителей, общества или других людей, которые ожидают от них определенного выбора образования. Это может привести к тому, что люди выбирают образование, которое не соответствует их интересам или способностям.

В-четвертых, быстро меняющийся мир и рынок труда могут сделать выбор образования еще более сложным. Некоторые профессии, которые сегодня востребованы, могут быть ненужными через несколько лет. Это может привести к тому, что люди выбирают образование, которое может оказаться устаревшим в будущем.

Алгоритм выбора направления обучения на основе интересов пользователя решает проблему выбора образовательной траектории на основе интересов пользователя.

Алгоритм будет выстроен на основе данных пользователей сети ВКОНТАКТЕ. В основу алгоритма закладывается гипотеза, что человек выбирает направление обучения на основе собственных интересов. В качестве собственных интересов отлично подходят группы, в которых состоит пользователь, его музыкальные предпочтения. Эти признаки наиболее точно могут определять интересы пользователя. Так как только пользователь принимает решение вступать в сообщество или нет, слушать эту музыку или нет.

Данная социальная сеть выбрана, за счет большой аудитории не только России, но и других стран. Общая концепция выявления выбора направления обучения представлена на рисунке 3.1.

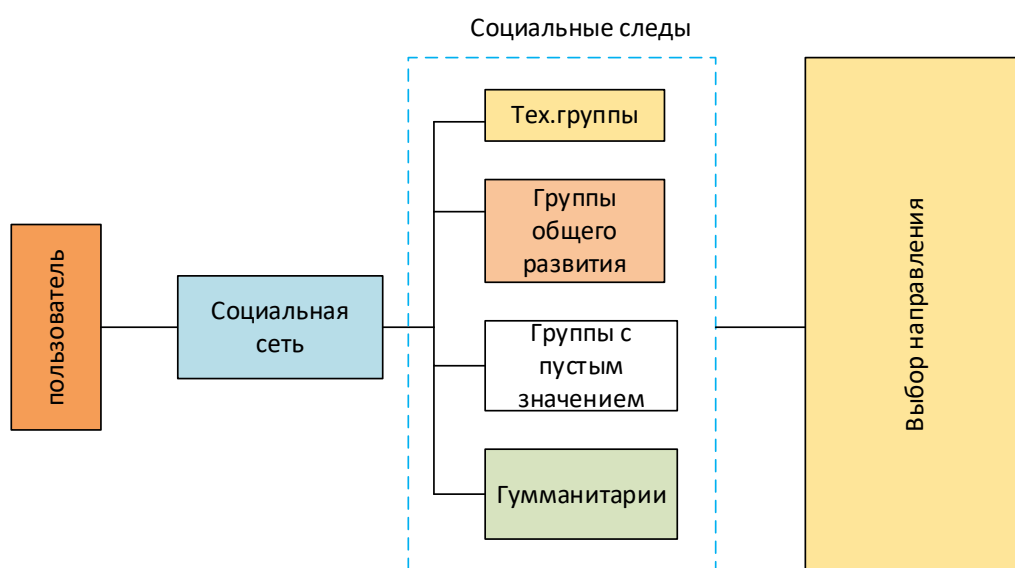


Рисунок 3.1 – Выявление предпочтений на основе следов в социальных сетях

На основе алгоритма будет созданы платформа рекомендаций по выбору направления обучения. Примерная концепция платформы представлены на рисунках 4.1а,б,в.

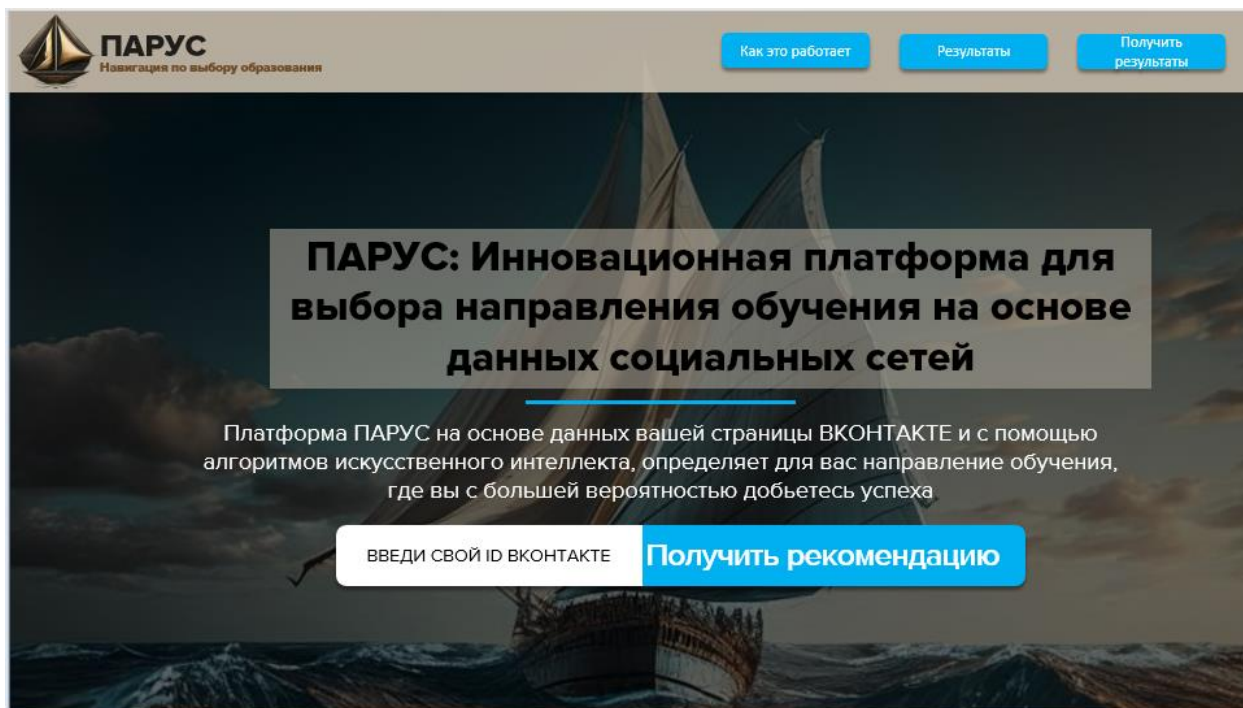


Рисунок 3.1а– Главная страница веб-приложения

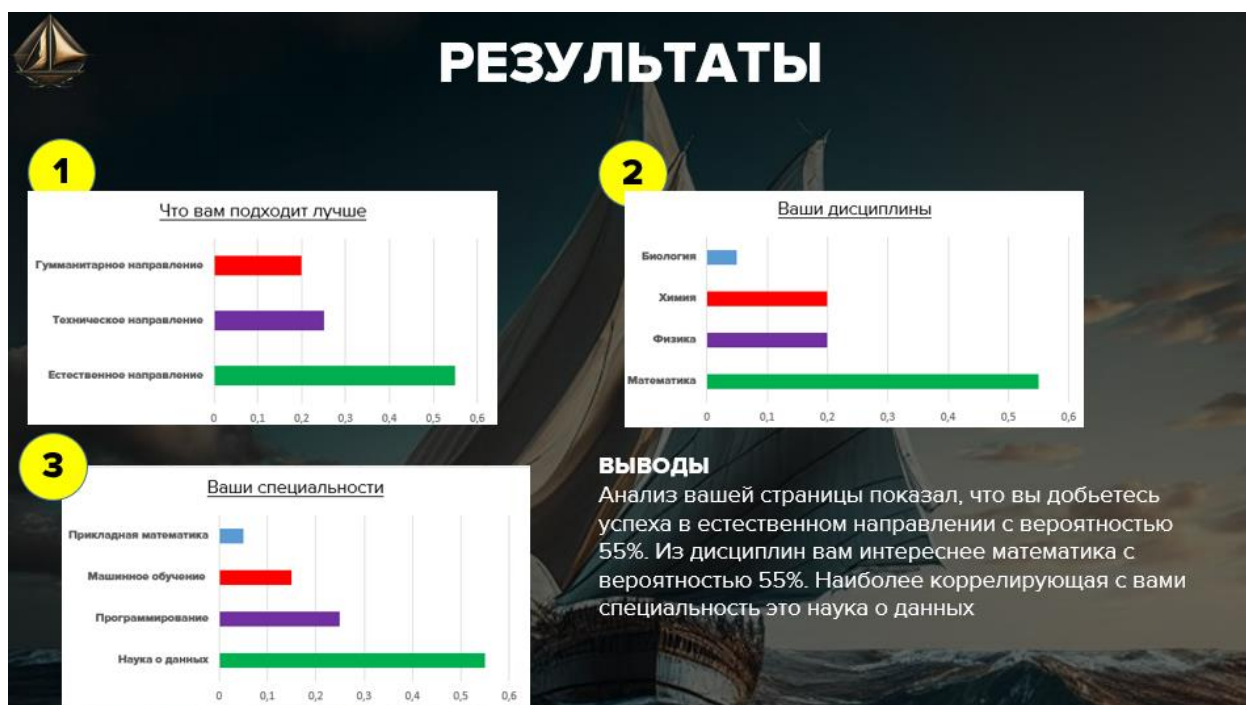


Рисунок 3.1б – Страница с результатами анализа

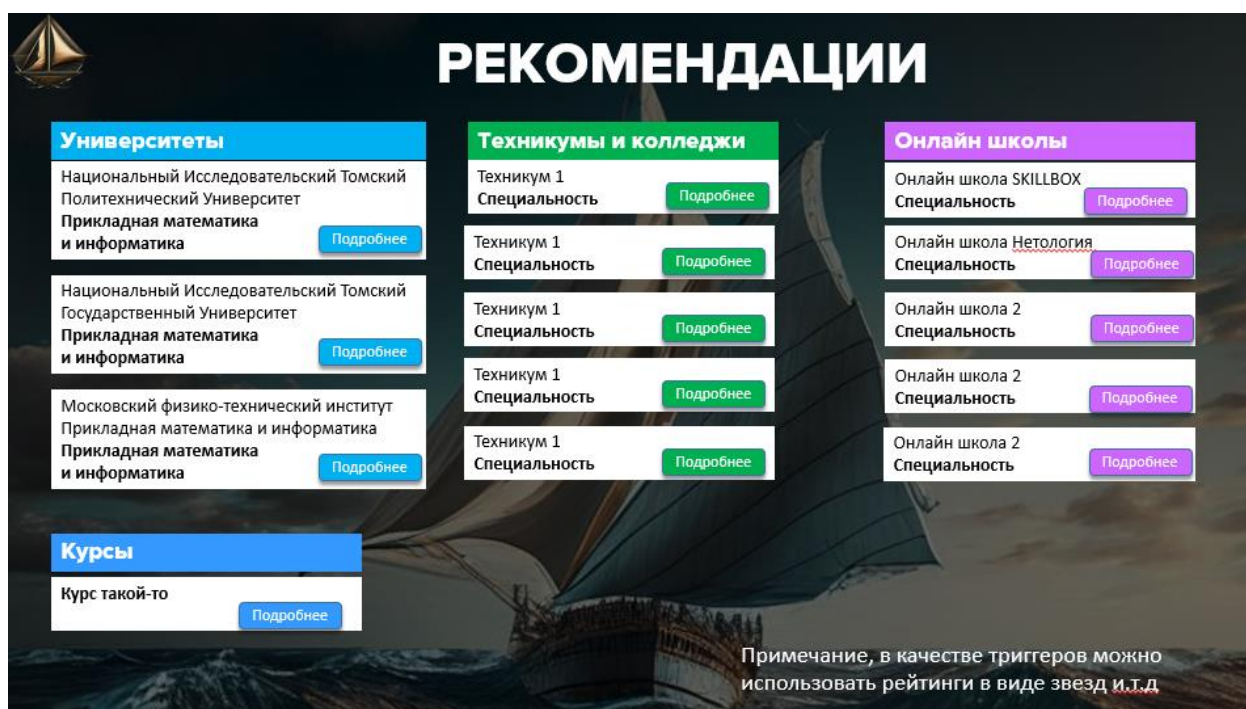


Рисунок 3.1в – Страница с рекомендациями курсов и учебных заведений

На данный момент тема выбора образования не является новой, при этом методология, которая сейчас задействована больше всего – это анкетирование. Анкетирование, не может показать полную картину интересов респондента, а признаки в социальной сети, довольно точно могут помочь раскрыть истинные интересы пользователей.

3.2 Объем и емкость рынка

Объем рынка для платформы, в основе которой лежит алгоритм выбора направления обучения в России может быть значительным, учитывая растущую популярность онлайн-образования и использование социальных сетей в повседневной жизни. Чтобы определить объем рынка необходимо базово описать бизнес-модель платформы с помощью чего платформа будет получать прибыль), рассчитать объем на основе данных пользователей сети вконтакте и рынка онлайн-образования.

Объем и емкость рынка будет рассчитана с учетом средних реферальных отчислениях онлайн-школ, которая составляет 16%.

Начальная бизнес-модель основывается на реферальных отчислениях, где мы получаем процент от оплаты за курсы и программы обучения, которые мы рекомендуем нашим пользователям. Схема базовой бизнес-модели представлен на рисунке 4.2

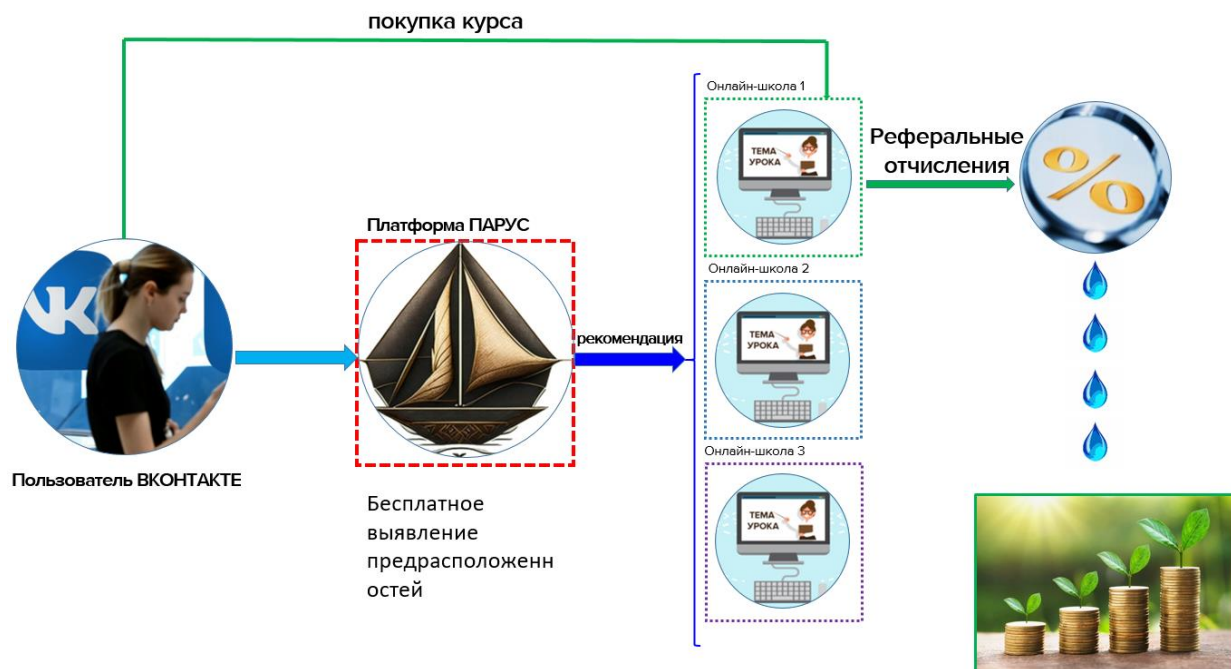


Рисунок 3.2 – Схема базовой бизнес модели

Ежемесячная аудитория в социальной сети ВКОНТАКТЕ за 4 квартал 2022 г представлена на рисунке 4.3.



Рисунок 3.3 – Ежемесячная аудитория в социальной сети ВКОНТАКТЕ [20]

Демографические данные по сети представлены на рисунке 3.4.



Рисунок 3.4 – Демография в ВКОНТАКТЕ [21]

Потенциальный объем рынка онлайн-образования для России.

Для расчета возьмем среднее население России 146 [22] млн. человек и объем платных образовательных услуг 4748 рублей за 2019 год [23].

Тогда РАМ составит

$$\begin{aligned} \text{РАМ}_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ}(2019)} &= M_{\text{ОБР.УСЛУГ}} \cdot \text{Население} \cdot K_{\text{РЕФ}} = \\ &= 4748 \cdot 146 \cdot 10^6 \cdot 0,16 = 110,9 \text{ млрд.руб,} \end{aligned}$$

где $K_{\text{РЕФ}} = 0,16$ – средние реферальные отчисления [24].

Учитывая ежегодный прирост суммы платных образовательных услуг на душу населения в пределах минимального среднегодового показателя в 6%, определим показатель стоимости услуг на 2023 год.

$$\text{СОУ}_{2023} = (1+k)^{\text{Текущий год}-2019} = 4748 \cdot (1+0,06)^{2023-2019} = 5994 \text{ руб}$$

Определим показатель РАМ за 2023 год

$$\begin{aligned} \text{РАМ}_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ}(2023)} &= \text{СОУ}_{2023} \cdot \text{Население} \cdot K_{\text{РЕФ}} = \\ &= 5994 \cdot 146 \cdot 10^6 \cdot 0,16 = 140 \text{ млрд.руб.} \end{aligned}$$

Общий целевой рынок ТАМ

Будем считать, что возрастная категория от 6 до 64 лет России является потенциальным клиентом. Общее население такой аудитории составляет

Население₆₋₆₄ = 113532429 человек

$TAM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ}(2023)} = COY \cdot \text{Население}_{6-64} \cdot K_{\text{РЕФ}} = 113532429 \cdot 5994 \cdot 0,16 = 109 \text{ млрд.руб.}$

Доступный объем рынка SAM

Чтобы рассчитать доступный объемы рынка уменьшим целевой возрастной диапазон потенциальной аудитории до 15-64 года. Общее количество населения этого возраста составляет:

Население₁₅₋₆₄ = 98111986 человек

Около 80% опрошенных не против пройти обучение с использованием онлайн-технологий.

Тогда SAM составит:

$SAM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ}(2023)} = 0,8 \cdot COY \cdot \text{Население}_{15-64} \cdot K_{\text{РЕФ}} =$
 $= 0,8 \cdot 98111986 \cdot 5994 \cdot 0,16 = 75,2 \text{ млрд.руб.}$

Реально достижимый объём рынка SOM

Условно средний процент охвата онлайн-технологиями составляет 49 % из них 30% готовы оплачивать онлайн обучение.

$SOM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ}(2023)} = 0,49 \cdot 0,3 \cdot COY \cdot \text{Население}_{15-64} \cdot K_{\text{РЕФ}} =$
 $= 0,49 \cdot 0,25 \cdot 98111986 \cdot 5994 \cdot 0,16 = 13,76 \text{ млрд.руб.}$

Учтем конкурентов агрегаторов онлайн-курсов.

Конкуренты представлены в таблице 4.1

Таблица 4.1 – Агрегаторы онлайн-курсов

Агрегатор	Посещаемость в год	Конверсия в переход, %	Конверсия в покупку, %	Средний чек	Реферальные отчисления, %	Выручка агрегатора, млн.руб
tutortop.ru	1975536	10,00%	4%	15500	16%	19,59731712
choosecourse.ru	198000	10,00%	4%	15500	16%	1,96416
skillscanner.io	396000	10,00%	4%	15500	16%	3,92832
obrazoval.ru	606000	10,00%	4%	15500	16%	6,01152
uchis-online.ru	1196400	10,00%	4%	15500	16%	11,868288
sravni.ru/kursy/	12000000	10,00%	4%	15500	16%	119,04
gdekurs.ru	274800	10,00%	4%	15500	16%	2,726016
go-course.ru	45600	10,00%	4%	15500	16%	0,452352
courselist.ru	1320	10,00%	4%	15500	16%	0,0130944
coursestop.ru	25200	10,00%	4%	15500	16%	0,249984
learningclub.ru	22800	10,00%	4%	15500	16%	0,226176
Общий годовой доход агрегаторов, млн.руб						166,07

Нужно отметить, что по данным аналитического отчета НЕТОЛОГИИ средняя цена курса за 2021 год составляла 15500 руб [25].

Тогда реально достижимый объем рынка с учетом конкурентов составит

$$SOM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ-КОНК(2023)}} = SOM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ-КОНК(2023)}} - \text{Выручка}_{\text{Агрегаторы}} = \\ = 13,76 - 166,07 \cdot 10^{-3} = 13,59 \text{ млрд.руб.}$$

При этом за три года, планируется что платформа займет 0,1% от реально достижимого объема рынка, тогда

$$SOM_{\text{ДОСТИЖИМЫЙ}} = SOM_{\text{ОНЛАЙН-ОБРАЗОВАНИЯ-КОНК(2023)}} \cdot K_{\text{РЫНКА}} = \\ = 13,59 \cdot 0,001 \cdot 1000 = 13,59 \text{ млн.руб,}$$

где $K_{\text{РЫНКА}} = 0,001$ – планируемая доля реально достижимого объема рынка, которую займет платформа.

Представим результаты расчета в виде диаграммы (рисунок 3.5)

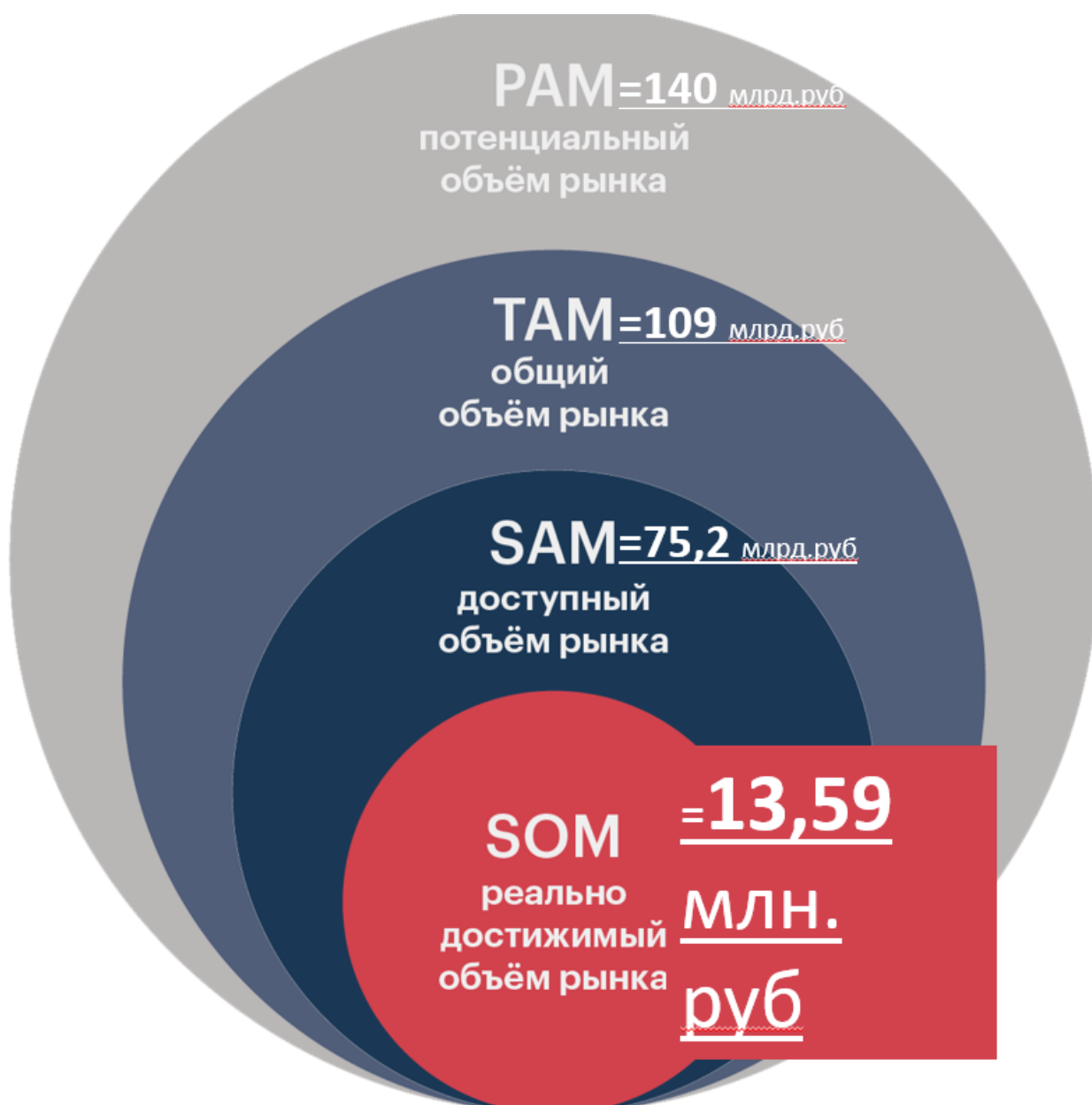


Рисунок 3.5 – Диаграмма объема рынка онлайн образования в России

3.3 Интеллектуальная собственность

Интеллектуальная собственность нужна для защиты интеллектуальных продуктов. Для получения свидетельства о регистрации программы для ЭВМ, проводится анализ патентов.

Найдем свидетельства о регистрации программ для ЭВМ подобной направленности. Свидетельства программ для ЭВМ подобной направленности представлен в таблице 3.5.

Таблица 3.5 – Программы для ЭВМ схожей направленности [26]

Номер регистрации (свидетельства)	Название программы	Реферат
2018617080	Комплекс тестирования «Профдиагностика»	Основное назначение программы - диагностика склонностей, способностей и личностных особенностей (черты темперамента) школьников 9-11 классов для подбора наиболее подходящих групп направлений обучения в профессиональных учреждениях, а также для выбора профиля обучения в старшей школе (для 9 классов).
2021661071	Оценка психофизического потенциала при выборе профиля обучения	Программа предназначена оценки психофизиологического статуса учащихся старших классов при выборе профиля обучения. Перед началом работы по карте интересов выбираются предпочтительные профили обучения. Для успешного обучения требуется соответствие свойств высшей нервной деятельности особенностям выбранного направления.
2015610704	Компьютеризированный комплекс тестирования «Профшкола»	Компьютеризированный комплекс направлен на диагностику мотивации, личностных качеств и интеллектуальных способностей школьников 8-11 классов. Задача комплекса - профорентация школьников, рассматривающих вопрос выбора профильного образования, подбора факультативных и элективных занятий, дополнительных развивающих программ в рамках школьного образования, а также выбирающих предметы, по которым они будут сдавать ЕГЭ (или ОГЭ).
2015610705	Компьютеризированный комплекс тестирования «Профвыбор»	Компьютеризированный комплекс совмещает анализ мотивационной сферы, личностных качеств и интеллектуальных способностей в рамках диагностики профессиональных склонностей. Основное назначение комплекса - профорентация школьников средних и старших классов (8-11 классы), рассматривающих вопрос выбора специальности или группы специальностей для дальнейшего обучения в системе профессионального образования.

Проведенный патентный анализ показал, что на территории РФ зарегистрировано около 10 программ для ЭВМ, реализованные на разных технологиях.

Главная техническая новизна моей разработки в том, что профессиональная ориентация подбирается через анализ интересов пользователя социальных сетей. Для защиты алгоритма от незаконного использования и распространения программа будет зарегистрирована в Роспатенте и получено свидетельство о регистрации программы для ЭВМ.

3.4 Анализ современного состояния и перспектив развития отрасли

В данном разделе будет приведена динамика развития онлайн-школ в России. Данное направление выбрано потому, что бизнес-модель платформы ПАРУС подразумевает получение, прибыли за счет комиссии от стоимости предлагаемых продуктов на рынке образования.

На данный момент, на рынке онлайн-образования в России наиболее крупные и известные платформы это Skillbox, Skyeng, Stepik, Яндекс Практикум, Нетология, Фоксфорд и другие. Они предлагают широкий выбор курсов и программ обучения в различных областях, от программирования и IT до бизнеса и маркетинга.

Кроме того, многие университеты и школы также начали активно развивать онлайн-образование, предлагая свои курсы и программы через собственные платформы или сотрудничая с уже существующими онлайн-образовательными сервисами.

В целом, рынок онлайн-образования в России продолжает расти и развиваться, предоставляя новые возможности для образования и профессионального роста. Рейтинг крупнейших edtech-компаний России за 2022 г. Представлен на рисунке 3.6.

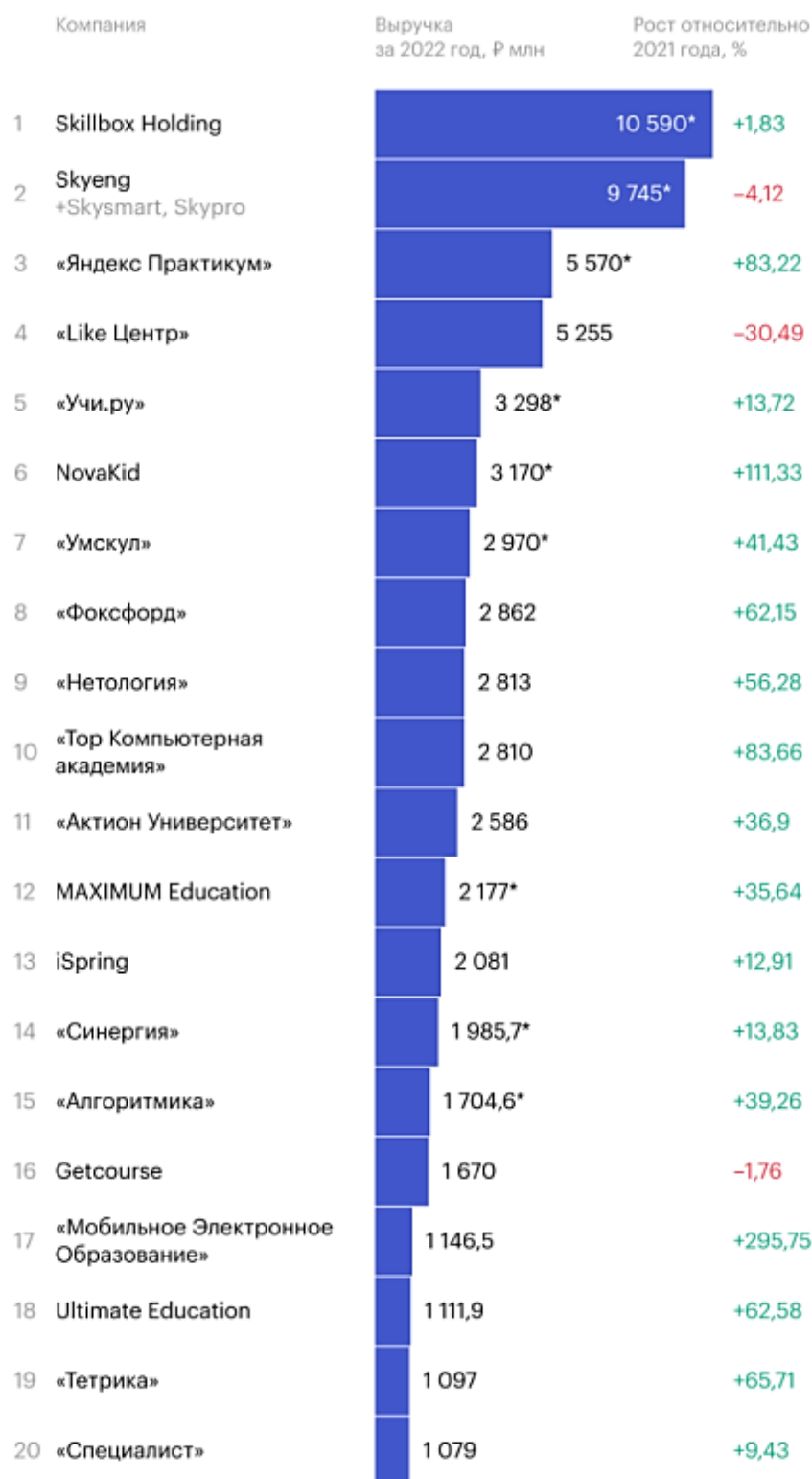


Рисунок 3.6 – Рейтинг крупнейших edtech-компаний России за 2022 г [27]

Как видно из данной диаграммы, онлайн школы продолжают свое развитие. На рисунке 4.7 представлен рост рынка EDTECH в мире.

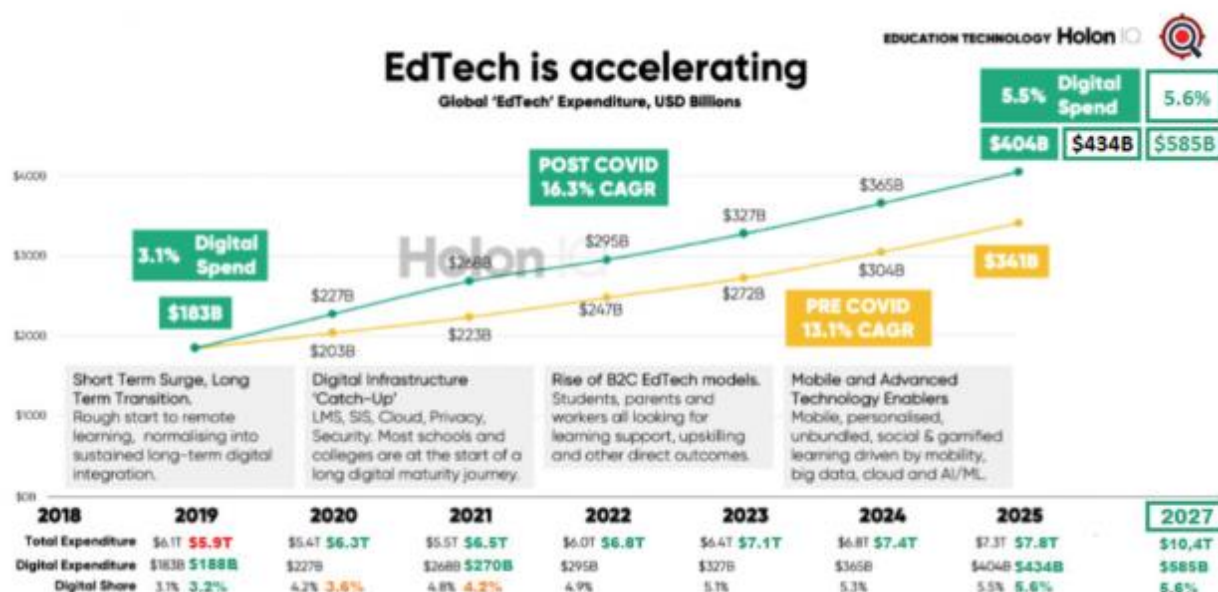


Рисунок 3.7– Рост рынка EDTECH в мире [28]

Как видно, из графика прогнозной динамики рынка EDTECH, То можно с уверенностью сказать, что рынок будет развиваться.

3.5 Планируемая стоимость продукта

Для монетизации платформы будут заключены партнерские соглашения с ведущими онлайн-школами, которые будут представлены в платформе ПАРУС. На платформе пользователям будут рекомендоваться курсы на основе анализа профиля в социальных сетях, определения интересов, предпочтений и потенциала. При регистрации на курс через нашу платформу, мы получим реферальные отчисления. Таким образом, мы будем зарабатывать на продажах курсов, а пользователи получают бесплатный доступ к персонализированным рекомендациям.

Также можно предоставлять дополнительные услуги для пользователей, такие как консультации с профессиональными консультантами.

Чтобы посчитать инвестиции на платформу разобьем этапы проектирования и запуска платформы на 6 этапов.

1. Исследование рынка и анализ конкурентов - 1-2 месяца
2. Разработка концепции и функциональных требований - 1 месяц

3. Дизайн и разработка пользовательского интерфейса – 1 -2 месяца
4. Разработка и тестирование функциональности платформы - 3-5 месяца
5. Запуск платформы и начало привлечения пользователей - 1 месяц
6. Оценка эффективности и оптимизация работы платформы - постоянно, в течение всего периода функционирования.

На данный момент готовы 1, 2 этапа находится на стадии завершения. Проведем расчет инвестиций для запуска платформы. Инвестиции на запуск платформы включают в себя постоянные расходы:

- заработная плата разработчика платформы;
- заработная плата специалиста по продвижению в социальных сетях;
- аренда сервера;
- покупка домена в зоне ru

Переменную часть:

- Затраты на рекламу

Стоит отметить, что разработка платформы будет осуществляться на локальном компьютере. Сервер будет арендован за месяц до запуска платформы. Это делается, для оптимизации трат на запуск платформы.

Определение трудоемкости выполнения работ и разработка графика проведения

При разработке платформы ПАРУС составляют трудовые затраты, поэтому определение трудоемкости проводимых работ является важным этапом составления бюджета.

Для определения ожидаемого (среднего) значения трудоемкости использована следующая формула:

$$t_{ожи} = \frac{3t_{мини} + 2t_{махи}}{5} ,$$

где

$t_{ожи}$ – ожидаемая трудоемкость выполнения i -ой работы, человеко-дни;

$t_{мини}$ – минимально возможная трудоемкость выполнения заданной i -ой работы, человеко-дни;

t_{maxi} – максимально возможная трудоемкость выполнения заданной i -ой работы, человеко-дни;

Для перевода длительности каждого этапа из рабочих в календарные дни, необходимо воспользоваться формулой:

$$T_{ki} = T_{Pi} \times k,$$

где

T_{ki} – продолжительность выполнения i -й работы в календарных днях;

T_{Pi} – продолжительность выполнения i -й работы в рабочих днях;

k – календарный коэффициент.

Календарный коэффициент определяется по формуле:

$$k = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{вых+праз}}} = \frac{365}{365 - 118} = 1,48,$$

где

$T_{\text{кал}} = 365$ – общее количество календарных дней в 2023 году;

$T_{\text{вых}}$ – общее количество выходных дней в году;

$T_{\text{праз+праз}} = 118$ – общее количество выходных/праздничных дней в году;

В таблице 4.6 показаны временные показатели для разработки

Таблица 3.6 – Временные показатели разработки платформы ПАРУС

Название этапа работы	Трудоемкость работ,			Длительность работ, дни		Должность исполнителя
	tmin	tmax	тож	Тр	Тк	
Регистрация ООО	3	5	4	4	6	Разработчик
Доработка функций по анализу социальных сетей	20	30	26	26	38	Разработчик
Дизайн и разработка пользовательского интерфейса	30	50	42	42	62	Разработчик
Разработка и тестирование функциональности платформы	90	140	120	120	178	Разработчик
Аренда и настройка сервера. Запуск Бета версии платформы	7	15	12	12	18	Разработчик
Разработка группы в социальных сетях и наполнение ее контентом	20	35	29	29	43	Специалист по продвижению в социальных сетях
Запуск платформы и начало привлечения пользователей	20	35	29	29	43	Специалист по продвижению в социальных сетях
Оценка эффективности и оптимизация работы платформы	59	59	59	59	87	Разработчик
Итого:	246	364	242	242	358	

Визуальное представление этапов разработки проекта (диаграмма Ганта) представлена на рисунке 3.8.

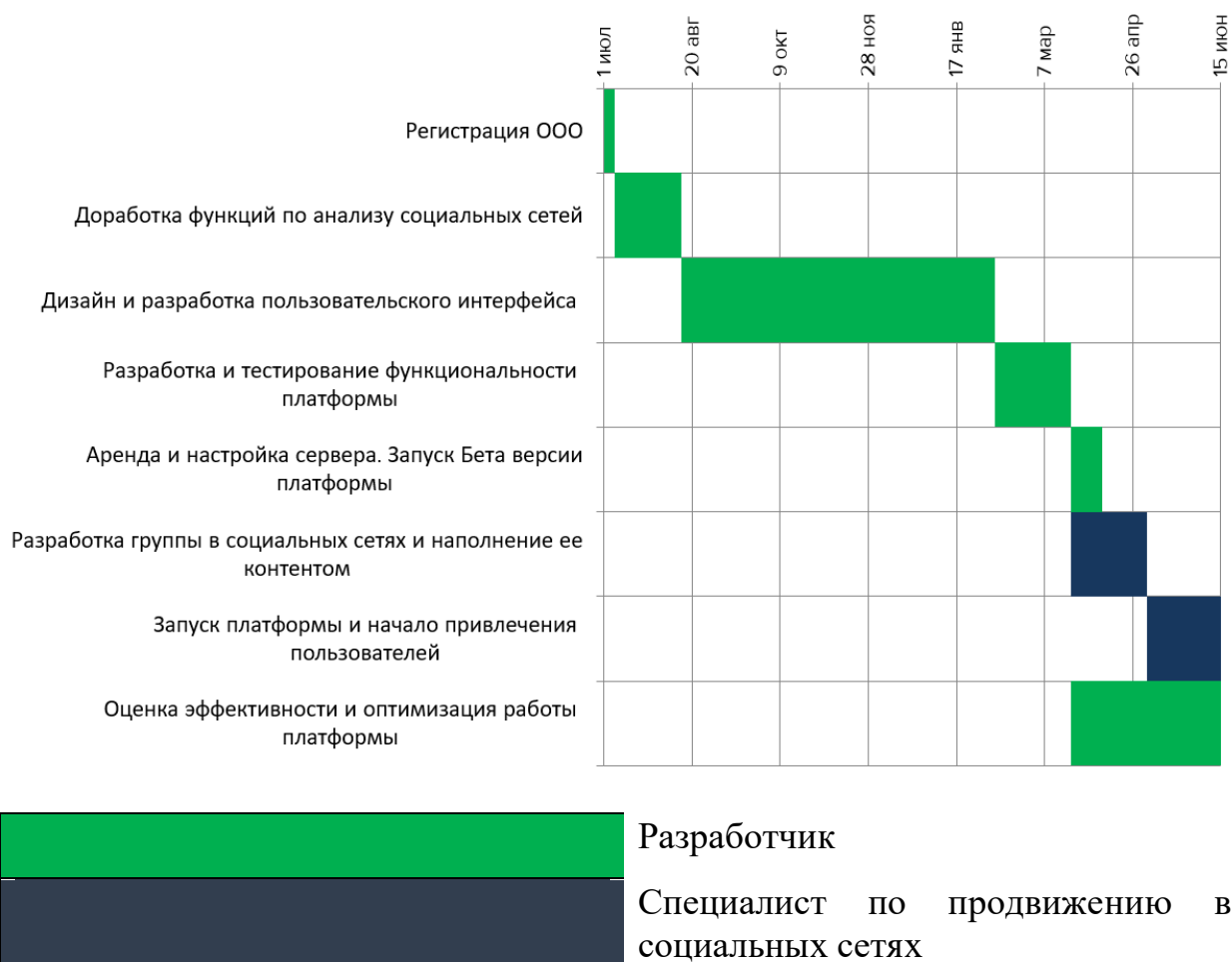


Рисунок 3.8 – Графическое представление этапов разработки платформы ПАРУС

Регистрация ООО

При регистрации ООО можно выделить следующие статьи расходов:

- оплата госпошлины — 4000 рублей;
- покупка юридического адреса — от 2000 рублей;
- внесение уставного капитала — минимум 10 000 рублей.

Тогда минимальные расходы на регистрацию ООО составят $I_{\text{рег.ooo}} = 16000$ руб.

Расчет амортизации специального оборудования

Расчет сводится к определению амортизационных отчислений, так как оборудование было приобретено до начала разработки и эксплуатировалось ранее, поэтому при расчете затрат на оборудовании учитываем только рабочие дни. Цена компьютера составляет 90000 руб.

Расчет амортизации проводится следующим образом:

Норма амортизации: рассчитывается по формуле:

$$A_H = \frac{1}{n}$$

где n – срок полезного использования в количестве лет.

Срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года.

$$A_H = \frac{1}{3} = 0,333$$

Годовые амортизационные оборудования разработчика:

$$И_{\text{ОСН.РАЗРАБОТЧИК}}^{\text{ОСН.РАЗРАБОТЧИК}}_{\text{АМ}} = \frac{T_{\text{РАБ}}}{T_{\text{КАЛ.ГОД}}} A_H \cdot C_{\text{КОМП}} = \frac{247}{365} \cdot 0,333 \cdot 90000 = 20281 \text{ руб},$$

где,

$T_{\text{РАБ}} = 247$ д. -число рабочих дней работы над разработкой (разработчика).

$T_{\text{КАЛ.ГОД}} = 365$ дня - действительный годовой фонд календарного времени.

Годовые амортизационные отчисления оборудования специалиста по продвижению в социальных сетях:

$$И_{\text{ОСН.СОЦСЕТИ}}^{\text{ОСН.СОЦСЕТИ}}_{\text{АМ}} = \frac{T_{\text{РАБ}}}{T_{\text{КАЛ.ГОД}}} A_H \cdot C_{\text{КОМП}} = \frac{58}{365} \cdot 0,333 \cdot 90000 = 4762 \text{ руб},$$

где,

$T_{\text{РАБ}} = 58$ д. -число рабочих дней специалиста по продвижению в социальных сетях.

$T_{\text{КАЛ.ГОД}} = 365$ дня - действительный годовой фонд календарного времени.

Итоговые амортизационные отчисления составят:

$$И_{\text{АМ}} = И_{\text{ОСН.СОЦСЕТИ}}^{\text{ОСН.СОЦСЕТИ}}_{\text{АМ}} + И_{\text{ОСН.РАЗРАБОТЧИК}}^{\text{ОСН.РАЗРАБОТЧИК}}_{\text{АМ}} = 4762 + 20281 = 25043 \text{ руб}.$$

Расчет заработной платы разработчика и специалиста по продвижению в социальных сетях.

Величина заработной платы определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и тарифных ставок. В состав основной заработной платы включается премия, выплачиваемая ежемесячно из фонда заработной платы в размере 20-30 % от тарифа или оклада.

Среднедневная заработная плата разработчика рассчитывается по формуле:

$$З_{\text{ДН(РАЗРАБОТЧИК)}} = \frac{З_{\text{М(РАЗРАБОТЧИК)}} \cdot М}{F_{\text{Д}}} = \frac{20000 \cdot 10,4}{247} = 842,1 \text{ руб.},$$

где

$З_{\text{М(РАЗРАБОТЧИК)}}$ = 20000 руб. оклад разработчика

$М=10,4$ мес – количество месяцев работы без отпуска в течение года;

$F_{\text{Д}}=247$ дня- годовой фонд рабочего времени в 2023 году;

Среднедневная заработная плата специалиста по продвижению в социальных сетях рассчитывается по формуле:

$$З_{\text{ДН(СОЦ.СЕТИ)}} = \frac{З_{\text{М(СОЦ.СЕТИ)}} \cdot М}{F_{\text{Д}}} = \frac{27000 \cdot 10,4}{247} = 1137 \text{ руб.},$$

где

$З_{\text{М(СОЦ.СЕТИ)}}$ = 27000 руб. оклад специалиста по продвижению в социальных сетях.

$М=10,4$ мес – количество месяцев работы без отпуска в течение года;

$F_{\text{Д}}=247$ дня- годовой фонд рабочего времени в 2023 году;

Разработчик после завершения работы занимается постоянной оптимизацией платформы, поэтому его рабочее время 247 дней.

При учете заработной платы специалиста по продвижению в социальных сетях оклад берется такой же как у разработчика. При расчете заработной платы учитывается премия в размере 20 %.

Затраты на заработную плату:

$$З_{\Pi} = З_{\text{осн}} + З_{\text{доп}}$$

$З_{\text{осн}}$ – основная заработная плата, руб.

$З_{\text{доп}}$ – дополнительная заработная плата, руб.

Заработная плата основная:

$$З_{\text{осн}} = З_{\text{дн}} \cdot T_p \cdot (1 + K_{\text{пр}} + K_{\text{д}}) \cdot K_p$$

$З_{\text{дн}}$ – среднедневная заработная плата, руб.

$K_{\text{пр}}$ – премиальный коэффициент (0,2- 0);

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

T_p – продолжительность работ, выполняемых работником, раб.дни

Расчет основной заработной платы представлен в таблице 3.7

Таблица 3.7 – Расчет основной заработной платы

Исполнители	Здн, руб.	Кпр	Кд	Кр	Тр	Зосн
Разработчик	842,1	0,2	0	1,3	247	324478,0
Специалист по продвижению в социальных сетях	1137,0	0,2	0	1,3	56	99328,3
Итого:						423806,3

Дополнительная заработная плата составляет 10% от основной заработной платы

Дополнительная заработная плата на разработку

$$З_{\text{доп}} = 0,1 \cdot З_{\text{осн}} = 0,1 \cdot 423806 = 42380,6 \text{ руб.}$$

Затраты на заработную плату

$$З_{\Pi} = З_{\text{осн}} + З_{\text{доп}} = 423806,3 + 42380,6 = 466187 \text{ руб.}$$

Расчет страховых взносов

Данная статья отражает обязательные отчисления по установленным законодательным нормам органам государственного социального

страхования, пенсионного фонда, государственного фонда занятости и медицинского страхования.

Затраты на страховые взносы рассчитываются как доля (30%) от затрат на оплату труда:

$$И_{\text{соц}} = 0,3 \cdot З_{\text{п}} = 0,3 \cdot 466187 = 139856 \text{ руб.}$$

Налог на доходы физических лиц

$$НДФЛ = 0,13 \cdot З_{\text{п}} = 0,13 \cdot 466187 = 60604,3$$

Расчет материальных затрат

В аренду будет взят выделенный сервер со следующей конфигурацией:

Процессор **Intel Xeon E-2386G 3.5 ГГц**, 6 ядер

Память 64 Гб DDR4

Диск 2 × 960 Гб SSD NVMe

Стоимость сервера согласно [] составляет 21000 руб/месяц

Для расчета затрат сервер будет взят на три месяца. До конца этапа запуска.

Тогда расчет на материальные затраты составит

$$И_{\text{сервер}} = n_{\text{мес}} \cdot Ц_{\text{сервера}} = 3 \cdot 21000 = 63000 \text{ руб.}$$

Переменные затраты (Затраты на рекламу)

Переменные затраты включают в себя затрату на рекламу.

Бюджет на рекламу составит 75000 руб/месяц. На начальном этапе реклама будет показываться только в социальной сети ВКОНТАКТЕ, так как функционал платформы на данный момент сосредоточен для работы с этой социальной сетью.

Реклама с оптимизацией будет запущена на 2 месяца.

Тогда затраты на рекламу составят

$$З_{РЕКЛАМА} = n_{МЕС} \cdot Б_{РЕКЛАМЫ} = 2 \cdot 75000 = 150000 \text{ руб.}$$

Накладные расходы

При разработке учитывают прочие затраты, не попавшие в предыдущие статьи расходов: оплата услуг связи, электроэнергии, почтовые расходы и т.д. Величина накладных расходов будет составлять 15% от общей суммы затрат. Их величина определяется по следующей формуле:

$$\begin{aligned} И_{НАК} &= 0,15 \cdot (И_{РЕГ.ООО} + И_{СЕРВЕР} + З_{П} + И_{СОЦ} + И_{АМ} + З_{РЕКЛАМА}) = \\ &= 0,15 \cdot (16000 + 63000 + 466187 + 139856 + 25043 + 150000) = 129013 \text{ руб.} \end{aligned}$$

В таблице 3.8 представлены общие затраты на разработку и запуск платформы ПАРУС

Таблица 3.8 – Общий бюджет на разработку платформы ПАРУС

Наименование	Сумма, руб.
Затраты на регистрацию ООО	16000
Затраты на аренду сервера	63000
Амортизация основных средств	25043
Затраты на дополнительную заработную плату	42380,6292
Накладные расходы	129013
Страховые взносы	139856,0764
Затраты на основную заработную плату	423806,292
Затраты на рекламу	150000
Общий бюджет	989098,9976

Стоит отметить, что при разработке платформы взят минимальный оклад, так как платформа будет разрабатываться самостоятельно, что поможет снизить затраты на разработку и ускорить процесс создания платформы.

4.6 Конкурентные преимущества создаваемого продукта, сравнение технико-экономических характеристик с отечественными и мировыми аналогами

Платформа ПАРУС – это абсолютно новый подход в рекомендациях обучения. Платформа позволяет выбирать пользователям наиболее

персонализированные курсы, чтобы каждый мог учиться в своем темпе и по своим интересам.

Одним из главных конкурентных преимуществ платформы являются персональные рекомендации на основе анализа профиля социальной сети.

Пользовательский интерфейс платформы достаточно простой и интуитивно понятный для большинства пользователей.

Как таковых платформ в мире еще нету, есть что-то похожее, поэтому чтобы сравнить конкурентные преимущества нужно понять как функционируют платформы:

Платформа Knewton

Персональные рекомендации пользователь платформы получает после диагностического теста. На основе результатов теста платформа создает индивидуальную образовательную программу, которая учитывает потребности и способности каждого ученика.

Платформа Knewton имеет несколько источников дохода. Они взимают плату за использование своей платформы учебными заведениями и учителями, которые используют Knewton для создания персональных учебных программ для своих студентов.

Кроме того, Knewton получает доход от продажи своих решений для онлайн-курсов и учебных материалов, которые могут быть интегрированы в другие платформы и сайты [29].

Платформа CollegeVine

Когда пользователь регистрируется на платформе CollegeVine, он может заполнить свой профиль, указав свои интересы, предпочтения и уровень образования. На основе этих данных и данных из социальных сетей, алгоритмы машинного обучения начинают анализировать информацию и предлагать наиболее подходящие курсы и материалы.

CollegeVine монетизирует свою платформу через платную подписку для студентов и их семей, которые ищут помощи в поступлении в университет.

Компания также получает комиссионные от университетов за каждого студента, который поступил в университет через CollegeVine [30].

Платформа Cialfo

Когда пользователь регистрируется на платформе Cialfo, он может создать свой профиль, указав свои академические достижения, интересы и предпочтения в отношении будущей карьеры. На основе этих данных алгоритмы машинного обучения начинают анализировать информацию и предлагать наиболее подходящие университеты и программы обучения.

Cialfo монетизирует свою платформу через платную подписку для школ и консультантов по поступлению, которые используют ее для работы с учениками и клиентами. Также платформа получает комиссионные от университетов за каждого студента, который поступил в университет через Cialfo.

Edurank

Edurank - это сервис, который предоставляет аналитические данные об учебных заведениях, включая информацию о рейтингах, репутации и социальном влиянии. Основной источник дохода для Edurank - это продажа доступа к своей платформе и аналитическим данным учебным заведениям, которые заинтересованы в повышении своей репутации и привлечении большего количества студентов. Кроме того, Edurank может получать доход от консультационных услуг и сотрудничества с другими компаниями в образовательной сфере.

В целом, монетизация Edurank основана на предоставлении ценной информации и аналитики для учебных заведений, которые готовы заплатить за такую экспертизу.

Площадки-агрегаторы

Площадки-агрегаторы, агрегируют все онлайн-курсы, информацию по ним, скидки и акции. Привлекают пользователей за счет поискового трафика.

На таких площадках нет никаких рекомендаций, пользователь сам сравнивает и выбирает онлайн курс.

В таблице 3.9 представлено сравнение конкурентных преимуществ

Таблица 3.9 – Сравнение конкурентных преимуществ

	Представлены ли в России?	Рейтинги курсов, сравнение цен	Продажа доступа к своим услугам	Подготовка к поступлению в университет	Аналитические данные об учебных заведениях	Персонализированные рекомендации на анализе социальных сетей	Использует алгоритмы машинного обучения	Простой и понятный интерфейс
ПАРУС	✓	✓	✗	✗	✓	✓	✓	✓
Edurank	✗	✓	✓	✗	✓	✗	✗	✗
Knewton	✗	✓	✓	✗	✗	✗	✓	✗
CollegeVine	✗	✓	✓	✓	✗	✗	✗	✗
Cialfo	✗	✓	✓	✓	✗	✗	✗	✗
Площадки-агрегаторы	✓	✓	✗	✗	✓	✗	✗	✗

Стоит отметить, что платформа ПАРУС не продает консультации, не помогает поступить в университет, она выявляет предпочтения пользователя и на основе этих предпочтений делает индивидуальные рекомендации. Возможно в будущем, стоит подключить некоторый функционал с международных площадок, представленных в таблице 3.1.

3.7 Целевые сегменты потребителей создаваемого продукта

Целевые сегменты потребителей создаваемого продукта включают в себя широкий круг пользователей, начиная от студентов и выпускников школ, заканчивая работающими людьми, которые хотят изменить свою профессиональную сферу деятельности. Кроме того, продукт может заинтересовать родителей, которые ищут рекомендации для своих детей, а также учебные заведения и компании, которые могут использовать платформу для анализа потенциальных кандидатов на вакансии.

Для выявления более полной целевой аудитории стоит посмотреть на портрет клиента, составленный в рамках исследования российского рынка онлайн-образования 2021 [25]. Портрет потребителя онлайн-образования представлен на рисунке 4.9.

Также стоит учесть, что основной аудиторией для выбора ВУЗа будут абитуриенты (выпускники школ)

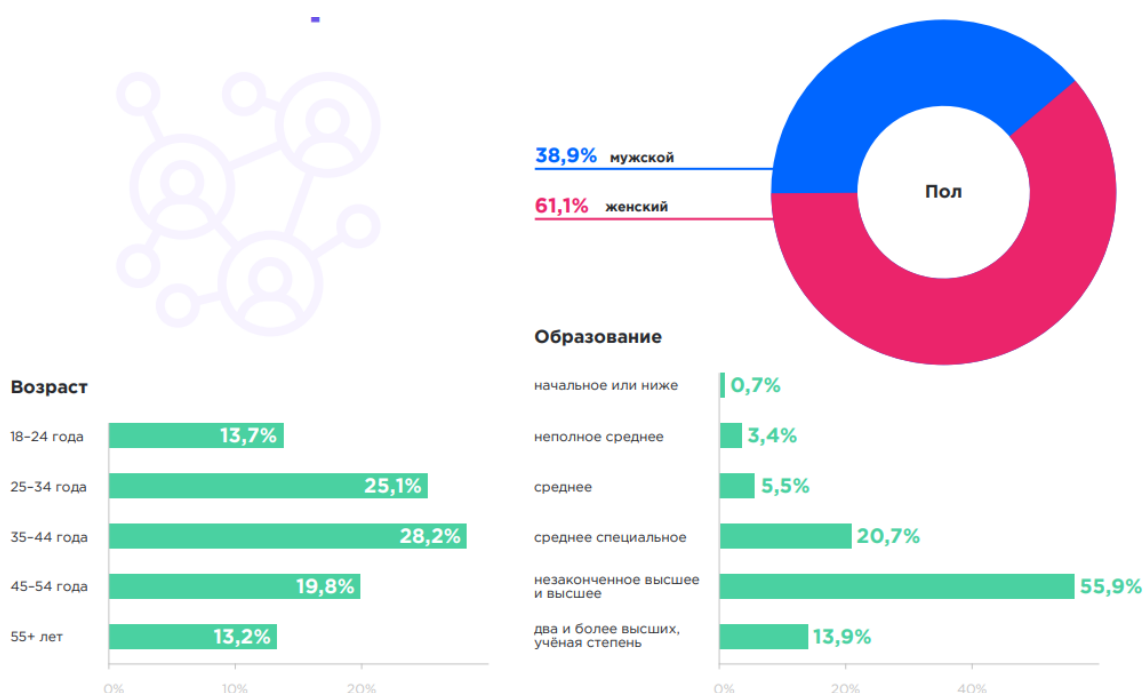


Рисунок 3.10 – Портрет потребителя онлайн-образования [25]

Целевая аудитория платформы ПАРУС может быть сегментирована на несколько групп в зависимости от их потребностей и интересов. Ниже приведены примеры таких групп:

Студенты: Студенты, которые хотят выбрать подходящее направление обучения и развить свою карьеру. Эта группа является наиболее важной для платформы ПАРУС, так как студенты активно ищут информацию о выборе направления обучения и могут стать активными пользователями платформы.

Молодые специалисты: Молодые специалисты, которые хотят развивать свою карьеру и получить новые знания и навыки. Эта группа также является важной для Платформы ПАРУС, так как молодые специалисты стремятся к профессиональному росту и могут быть заинтересованы в использовании платформы для выбора подходящих курсов обучения.

Специалисты среднего возраста: Специалисты среднего возраста, которые хотят обновить свои знания и навыки или изменить направление своей карьеры.

Учебные заведения, которые могут использовать платформу для анализа потенциальных студентов и улучшения процесса подбора.

Компании, которые ищут кандидатов на вакансии и могут использовать платформу для анализа потенциальных кандидатов и улучшения процесса подбора персонала. Особенно если, выбранное направление обучения, коррелирует с законченным образованием на эту тему.

Наиболее платежеспособными клиентам являются работающие люди, а значит молодые специалисты и специалисты среднего возраста. Также такое предположение основано на портрете потребителя онлайн-образования.

С точки зрения поступления в ВУЗ наиболее важными являются студенты и абитуриенты.

3.7 Бизнес-модели проекта. Производственный план и план продаж

В современном мире, где конкуренция на рынке труда и образования становится все более ожесточенной, необходимо иметь эффективный инструмент для выбора профессиональной сферы деятельности и развития своих навыков. Именно таким инструментом может стать платформа ПАРУС, которая помогает студентам и работающим людям получить персональные рекомендации для выбора направления обучения, чтобы сделать свою работу частью своих интересов и предпочтений.

Основной ценностью платформы ПАРУС является помощь студентам и работающим людям в выборе профессиональной сферы деятельности и развитии своих навыков. Поэтому, бизнес-модель должна быть ориентирована на улучшение качества продукта и удовлетворение потребностей целевых сегментов. Для достижения этой цели, необходимо разработать алгоритм, который обеспечит точность и полезность рекомендаций на платформе. Также, стоит уделять внимание стратегиям привлечения новых пользователей и укрепления имиджа платформы на рынке [34, 35].

Наиболее целесообразно для описания бизнес-модели использовать инструменты стратегического планирования – бизнес-модель Остервальдера. Данная бизнес–модель представлена на рисунке 3.10

Таблица 3.10 – Бизнес-модель Остервальдера [33]

 Ключевые партнеры - Онлайн-школы - Образовательные учреждения	 Ключевые виды деятельности Анализ социальных сетей и предоставление персонализированных рекомендаций	 Ценностное предложение Персонализированные рекомендации для выбора направления обучения на основе анализа профиля пользователя в социальной сети ВКОНТАКТЕ	 Отношения с клиентами - Самообслуживание через платформу путем ввода своего профиля ВКОНТАКТЕ - Поддержка e-mail коммуникации - Онлайн консультант	 Сегменты клиентов - Студенты - Молодые специалисты - Специалисты среднего возраста - Группы ВКОНТАКТЕ партнеров
 Ключевые издержки - Разработка и поддержка онлайн-платформы - Аренда сервера - Заработная плата. Страховые отчисления. Налоги.	 ДОХОДЫ Получение комиссии с покупки курсов, с рекомендаций платформы Кроме того, платформа ПАРУС может использовать другие способы формирования доходов, такие как продажа рекламы или предоставление дополнительных платных услуг.			

Рассчитаем доходность этой модели.

Для расчета доходности также необходимо рассмотреть воронку продаж. Воронка продаж показана с учетом того, как будет работать готовая платформа.



Рисунок 4.11 – Воронка продаж

Чтобы приблизительно оценить доход, необходимо оценить примерный охват пользователей, ищущих курсы или обучение. Для этого воспользуемся сервисом wordstat.yandex.ru.

Приведу пример для слова «Онлайн-Курсы» количество ежемесячных запросов составляет 375 535. Общая динамика запросов представлена на рисунке 4.12.

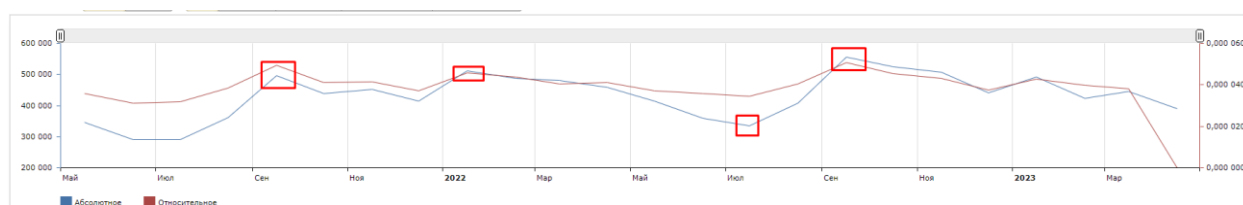


Рисунок 3.12 – Динамика запросов онлайн-школа

Остальные запросы будут представлены в таблице 3.12

Таблица 3.11 – Запросы в месяц на обучение

Запрос	Показов в месяц
skillbox курсы	8779
skillbox отзывы	5596
skillbox обучение	1421
skillbox python	1313
skyeng языки	1505
яндекс практикум	96948
курсы яндекс практикум	8587
яндекс практикум python	3606
яндекс практикум разработчик	2778
яндекс практикум обучение	2017
яндекс практикум аналитик	2141
яндекс практикум аналитика	2141
аналитика данных яндекс практикум	1283
яндекс практикум питон	873
школа фоксфорд	10433
онлайн школа фоксфорд	7348
фоксфорд курсы	3029
нетология курсы	6102
нетология отзывы	5909
промокод нетология	1155
нетология официальный сайт	10231
компьютерная академия тор	4151
компьютерная академия тор отзывы	671
синергия обучение	9934
синергия дистанционное обучение	2646
синергия стоимость обучения	2569
онлайн школа	375535
обучение программированию	26165
ИТОГО	604866

Это не все запросы, их можно набрать очень много, как целевых, так и коммерческих запросов.

Рассчитаем ежемесячный доход с платформы

Будем считать, что 70 % людей от суммы запросов. Присутствуют в социальной сети ВКОНТАКТЕ. Тогда условное количество людей, которые ищут для себя возможность для обучения составит:

$$\text{Тогда } N_{\text{ВК}} = M_{\text{ЗАПРОСОВ}} \cdot 0,7 = 0,7 \cdot 604866 = 423406 \text{ человека}$$

При этом планируется ежемесячно охватывать около 6%. Это означает, что 6% будет переходить на платформу. Мотивация посещающих следующая: посмотреть какая специальность подойдет лучше всего. Так как платформа будет ежемесячно улучшать, то и охват будет на 5-8% от планируемого охвата.

Тогда количество людей, кто перейдет на платформу составит:

$$N_{\text{ЛЮДЕЙ.ПЛАТФ}} = M_{\text{ВК}} \cdot 0,06 = 0,7 \cdot 423406 = 25404 \text{ человека.}$$

После этого люди на платформе останутся или просто посмотреть, что для них лучше или посмотрят рекомендации и перейдут на сайт с онлайн-курсом. Средняя конверсия в переход, согласуется с конверсией сайтов агрегаторов и составляет 10%, также необходимо учесть, что конверсия постепенно будет увеличиваться, за счет отслеживания действий пользователя и улучшения пользовательского интерфейса.

Тогда количество людей с платформу, которое перейдет на сайт онлайн-курса составит:

$$N_{\text{ОНЛАЙН.КУРС}} = N_{\text{ЛЮДЕЙ.ПЛАТФ}} \cdot 0,1 = 25404 \cdot 0,1 = 2540 \text{ человек.}$$

Средняя конверсия в покупку, кто перешел на страницу курса составляет 2,2%, тогда число покупок курса составит:

$$N_{\text{ПОКУПОК}} = N_{\text{ОНЛАЙН.КУРС}} \cdot C_{\text{ПОКУПКИ}} = 2540 \cdot 0,022 = 56.$$

Доход с покупки при средних реферальных отчислениях 16% и средней стоимости курса 15500 руб. составит:

$$\text{Доход} = N_{\text{ПОКУПОК}} \cdot \text{Цена}_{\text{КУРСА}} \cdot K_{\text{РЕФ}} = 56 \cdot 15500 \cdot 0,16 = 138800 \frac{\text{руб}}{\text{мес}}.$$

Финансовые потоки поквартально представлены в таблице 3.12

Таблица 3.12 – Финансовые потоки с платформы ПАРУС поквартально

Квартал с начала действия гранта	1	2	3	4	5	6	7	8	9	10
РАСХОДЫ										
Регистрация ООО	16000	0	0	0	0	0	0	0	0	0
Оклад разработчика, тыс.руб	20000	20000	20000	20000	20000	20000	20000	20000	20000	20000
Оклад обработчика заявок	0	0	0	0	0	0	0	0	0	0
Оклад специалиста по продвижению в социальных сетях	0	0	0	27000	27000	27000	27000	27000	27000	27000
Коэффициент премии	20%	20%	20%	20%	20%	20%	20%	20%	20%	20%
Районный коэффициент	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%
Зарплата до вычета НДФЛ	93600	93600	93600	219960	219960	219960	219960	219960	219960	219960
Дополнительная зарплата	9360	9360	9360	21996	21996	21996	21996	21996	21996	21996
Затраты на заработную плату	102960	102960	102960	241956	241956	241956	241956	241956	241956	241956
Социальные отчисления	30888	30888	30888	72586,8	72586,8	72586,8	72586,8	72586,8	72586,8	72586,8
Бюджет на рекламу	0	0	0	150000	30000	30000	30000	30000	30000	30000
Аренда сервера	0	0	21000	63000	63000	63000	63000	63000	63000	63000
Амортизационные отчисления, руб	6260,7	6260,7	6260,7	6260,7	6260,7	6260,7	6260,7	6260,7	6260,7	6260,7
Накладные расходы	23416,30 5	21016,30 5	21016,30 5	48120,52 5	48120,52 5	48120,52 5	48120,52 5	48120,52 5	48120,53 5	48120,52 5
Суммарные издержки	179525,0 05	161125,0 05	182125,0 05	581924,0 25	461924,0 25	461924,0 25	461924,0 25	461924,0 3	461924 3	461924,0 25
ДОХОДЫ										
Количество человек	423406	423406	423406	423406	423406	423406	423406	423406	423406	423406
Охват, %	0%	0%	0%	12%	21%	26%	33%	41%	53%	67%
Охват человек	0	0	0	52079	87493	110241	138903	175018	223148	284514
Конверсия в переход	0	0	0	10,10%	10,73%	11,39%	12,10%	12,85%	13,64%	14,52%

Число человек на сайте курса	0	0	0	5260	9385	12558	16804	22485	30446	41304
Конверсия в покупку (минимум)	0	0	0	2,20%	2,20%	2,20%	2,20%	2,20%	2,20%	2,20%
Средняя цена за курс , руб	15500	15500	15500	15500	15500	15500	15500	15500	15500	15500
Ставка отчислений комиссии	16%	16%	16%	16%	16%	16%	16%	16%	16%	16%
Комиссионные отчисления	2480	2480	2480	2480	2480	2480	2480	2480	2480	2480
Число покупок	0	0	0	116	206	276	370	495	670	909
Выручка	0	0	0	287680	510880	684480	917600	1227600	1661600	2254320
ПРИБЫЛЬ										
Прибыль	-179525,0	-161125,0	-182125,0	-294244,0	48956,0	222556,0	455676,0	765676,0	1199676,0	1792396,0
Налог на прибыль	0,0	0,0	0,0	0,0	9791,2	44511,2	91135,2	153135,2	239935,2	358479,2
Чистая прибыль	-179525,0	-161125,0	-182125,0	-294244,0	48956,0	222556,0	455676,0	765676,0	1199676,0	1792396,0
Инвестиции	1000000,0									
	-	-	-	-	-	-	-	-		
Куммулятивный доход	1179525,0	1340650,0	1522775,0	1817019,0	1768063,1	1545507,1	1089831,1	-324155,1	875520,8	2667916,8
Остаток инвестиций	820475,0	659350,0	477225,0	182981,0	231936,9	454492,9	910168,9	1675844,9	2875520,8	4667916,8

График остатка инвестиций поквартально представлен на рисунке 4.13



Рисунок 3.13 – График остатка Гранта поквартально с учетом прибыли

График вычисления срока окупаемости представлен на рисунке 3.14

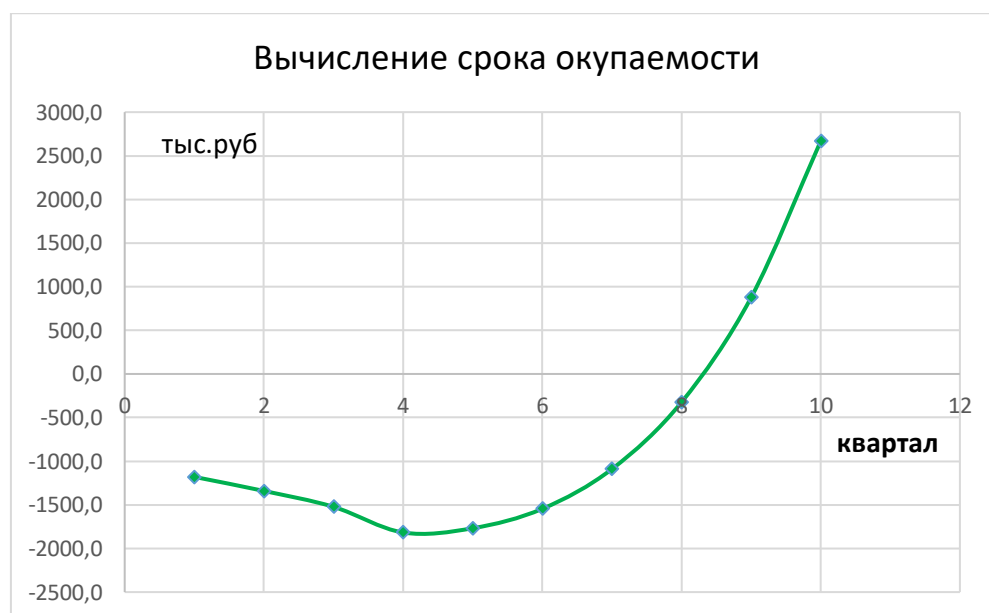


Рисунок 3.14 – Вычисление срока окупаемости (период времени квартал)

Из представленных выше расчетов можно сделать вывод, что проект окупится примерно через 25 месяцев после получения гранта. Через 30 месяце ожидаемый доход с платформы составит 1 млн 792 тыс.руб в месяц. Также стоит отметить, что первичное получение дохода начнется сразу после запуска платформы, то есть предприятие не уйдет в минус, что видно из графика на рисунке 4.14.

Члены команды

Все члены команды имеют опыт и компетенции в выбранных областях.

Команда представлена на рисунке 4.15.

Члены команды:		Ключевые роли:
	Гузеев Евгений Викторович, генеральный директор Магистратура по специальности «Компьютерные вычисления и математическое моделирование». - Опыт продвижения высокотехнологичных продуктов («Видеоаналитика по маркетингу») - Опыт в маркетинге и продажах около 2х лет.	Идеолог платформы. Разработка функционала платформы. Backend и Frontend разработка.
	Гузеева Анна Михайловна, главный бухгалтер Образование высшее экономическое. Опыт работы в бухгалтерии более 5 лет.	Учет финансовой деятельности предприятия. Работа по договору ГПХ
	Мальцев Алексей Викторович, Специалист по продвижению в социальных сетях Образование высшее техническое. Опыт в маркетинге и продвижении в социальных сетях 12 лет.	Продвижение платформы в социальных сетях. Привлечение клиентов
	Осинский Кирилл Юрьевич, инженер-программист Опыт в разработке более 10 лет. Основной стек: Java 11, Kotlin, Spring Boot, Kafka, PostgreSQL, Phoenix, Oracle	Разработка высоконагруженных частей платформы Работа по договору ГПХ

Рисунок 3.15 – Члены команды

Для минимизации затрат, основную часть работы над платформой будет реализовывать генеральный директор(разработчик). Примерно через 9 месяцев после начала разработки в работу включается специалист по продвижению в социальных сетях.

Главный бухгалтер и инженер программист работают по договору ГПХ до самоокупаемости проекта. Это сделано для оптимизации бюджета.

В будущем сотрудники будут набираться из студентов и выпускников профильных ВУЗов.

3.8 Стратегия продвижения продукта на рынок

Всю стратегию продвижения платформы на рынок условно можно поделить на платные и бесплатные методы. В общем виде стратегия продвижения представлена на рисунке 3.15

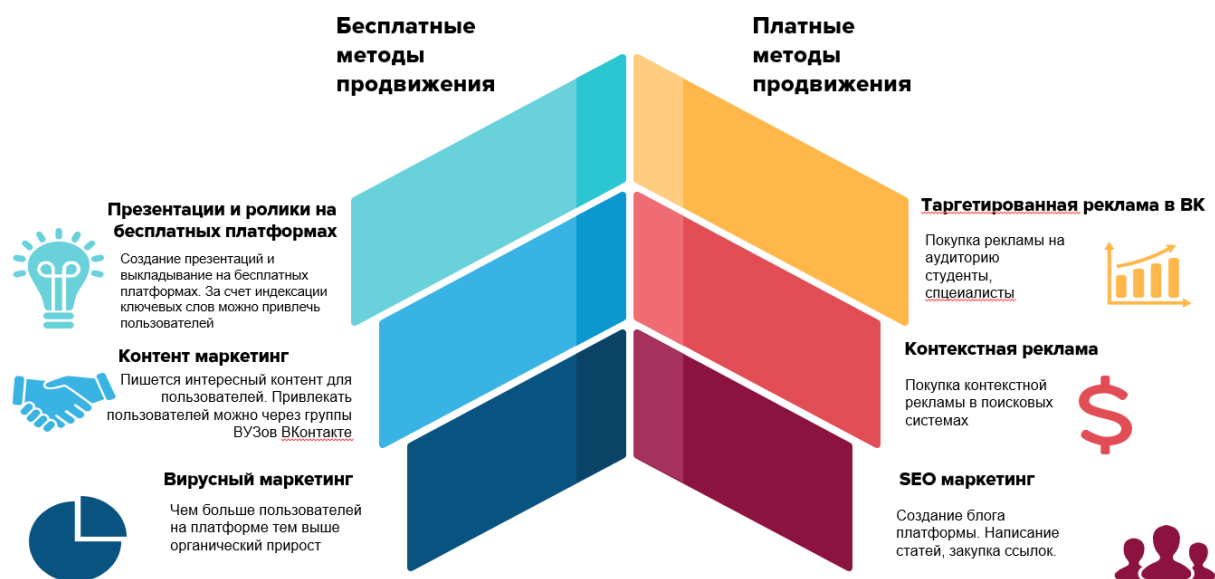


Рисунок 3.16 – Стратегия продвижения для платформы ПАРУС

Распишем эти методы подробнее:

БЕСПЛАТНЫЕ МЕТОДЫ:

Презентации и ролики на бесплатных платформах

Существуют такие платформы для презентаций, которые хорошо индексируются поисковыми системами. Одни из таких платформ это **prezi.com** (конструктор презентаций). Также агрегаторы презентаций: **pptcloud.net**, **myshared.ru**. Помимо этих платформ можно сделать презентационный ролик с ключевыми словами (низкочастотные ключевые слова). Также такой ролик можно выложить на видеоплатформы: **rutube.ru**, **youtube.ru**, **ВКОНТАКТЕ-видео**.

Схема работы по привлечению работает следующим способом:

1. Пользователь ищет что-то в сети, например, «тесты, для точного выбора образования». Если запрос низкочастотный, то с большой вероятностью ему появится ссылка на презентацию платформы.
2. На презентации платформы, будут указаны контакты, группа ВКОНТАКТЕ или сайт с платформы. Человек переходит и узнает о платформе.

Контент-маркетинг

Контент-маркетинг создается для вовлечения пользователя и увеличению лояльности пользователя. Весь контент-маркетинг будет вестись в группе ВКОНТАКТЕ о платформе и дублироваться в блоге.

Некоторые примеры контент плана для продвижения платформы ПАРУС представлены в таблице 3.13.

Таблица 3.13 – Некоторые темы для контент-маркетинга

Тема	Объяснение
Выбор образования	Советы, как выбрать образование согласно со своими интересами
Истории успеха	Истории успеха, как правильный выбор образования помогают достижению успеха в карьере
Общие тренды в образовании	Рассказывается о образовании, что будет актуально и через несколько лет
Обзоры онлайн-школ	Обзоры школ, которые будут представлены на платформе
Советы для эффективного обучения	Как эффективно учиться и все успевать
Карьерные треки	Показывать возможности в карьере различных областях
Интервью с экспертами	Приводятся интервью с экспертами в области профпригодности, HR-менеджерами
Мифы образования	разбор основных мифов образования Например «Высшую математику успешно на 1м семестре сдают только 10% студентов»
Стипендии и гранты	Какие на данный момент есть гранты, господдержка в обучении и.т.д.

Также нужно отметить, что очень много студентов сидят в группах ВУЗов, так как платформа бесплатна и может помочь как абитуриентам, так и студентам в выборе специализации. То можно договариваться с такими группами, они получают полезный контент, платформа получит внимание пользователей.

Вирусный маркетинг

По факту, вся информация в сети передается между пользователями. Если эта информация полезная, цепляющая или помогающая. Поэтому для вовлечения помимо контента, можно включать процессы игровой динамики, розыгрыши книг или абонементов на образовательные материалы. Такой вид маркетинга может быть, как платным так и бесплатным.

ПЛАТНЫЕ МЕТОДЫ ПРИВЛЕЧЕНИЯ:

Таргетированная реклама вконтакте – это реклама, которая будет показана пользователям. Сложность такой рекламы, в умении ее правильно настроить в условиях ограниченного бюджета. Общая схема запуска таргетированной рекламы представлена на рисунке 3.17



Рисунок 4.17 – Общая схема запуска рекламы

Контекстная реклама

Контекстная реклама, отображается в поисковых сетях, когда пользователь введет запрос. Для ее создания и минимизации трат, следует внимательно подбирать ключевые запросы и делать объявления кликабельными.

Пример такой рекламы представлен на рисунке 4.18

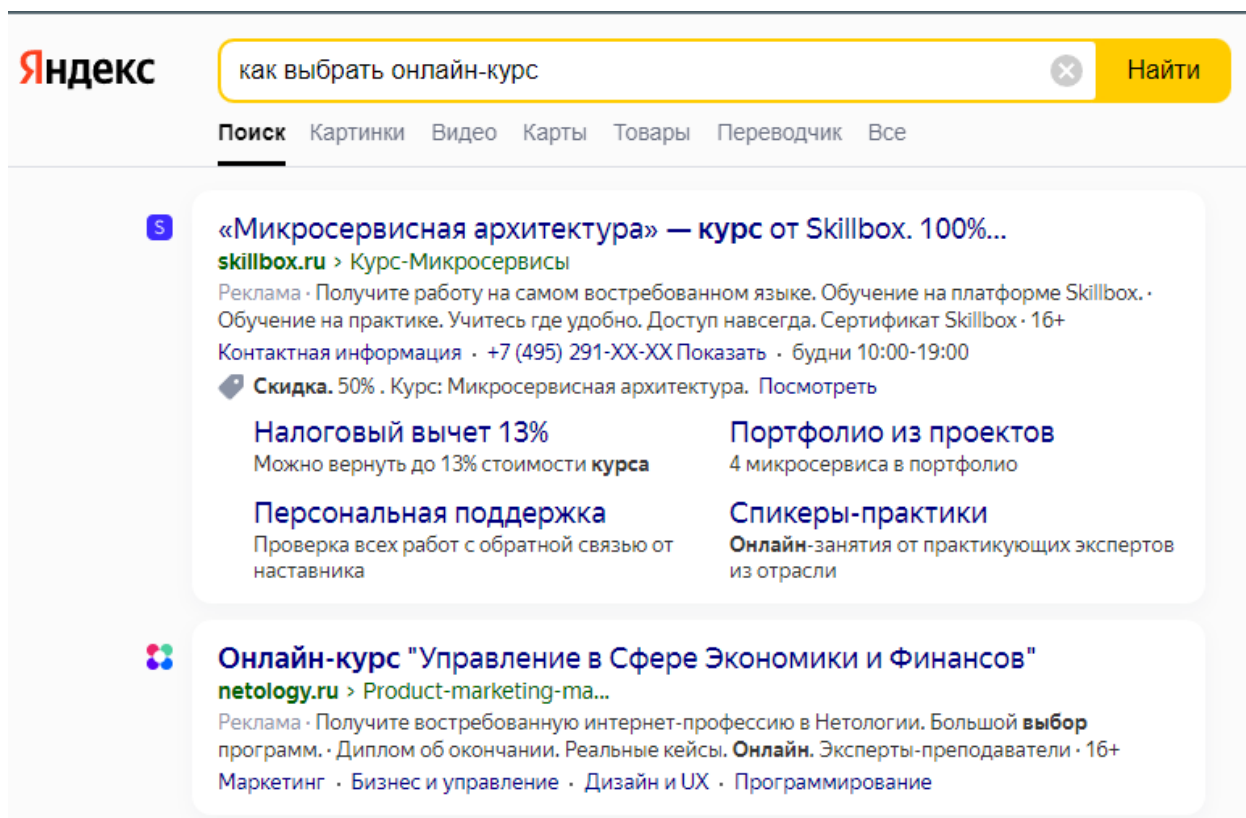


Рисунок 3.18 – Пример контекстной рекламы в Яндекс

Из рисунка 3.18 можно сделать вывод, что данные ключевые слова в рекламе будут достаточно дорогие. При этом рекламируются сразу онлайн-школы. У них высокие рекламные бюджеты, они могут настраиваться сразу на всю Россию. Поэтому очень важно подбирать ключевые слова и отсеивать ненужные (оптимизация трат).

SEO-маркетинг

Для SEO-маркетинга будет сделан блог, который будет вести на платформу. SEO-маркетинг это процесс оптимизации сайта, для улучшения его

видимости в поисковых системах. Это достаточно сложно и долго, но эффект от SEO существенный.

Общая схема SEO-маркетинга для платформы представлена на рисунке 3.19

SEO-маркетинг для платформы ПАРУС

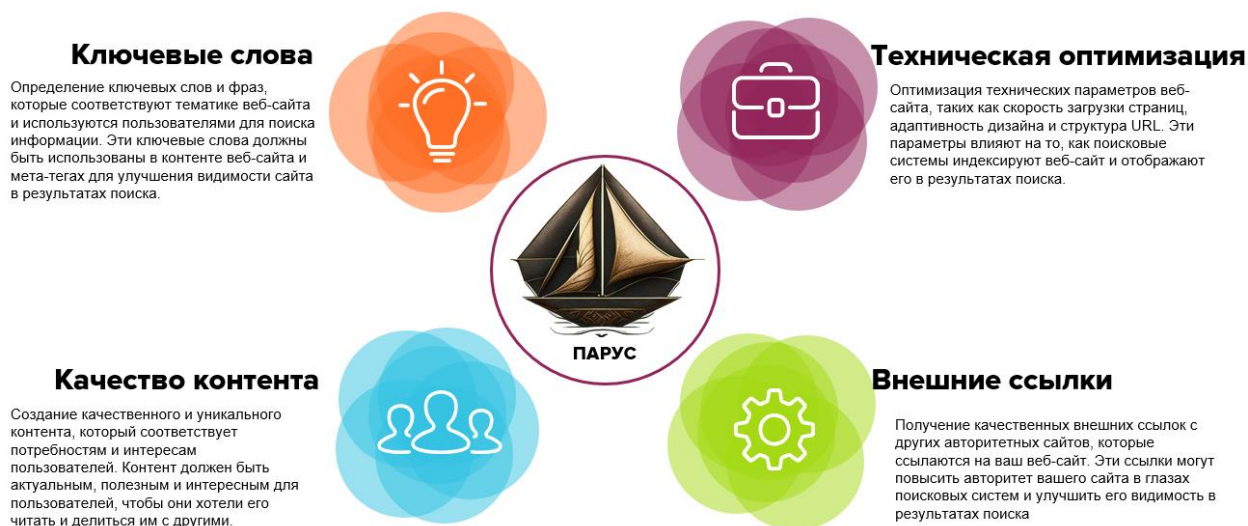


Рисунок 3.19 – Схема SEO-маркетинга для продвижения платформы

Как видно, что платформа будет продвигаться полностью онлайн из методов оффлайн можно участвовать в различных конференциях, выставках.

Для продвижения от начала к концу будут использованы следующие методы:

- создание группы в вконтакте, сайта, блога и контент маркетинг, работа с группами учебных заведений
- размещение информации на презентационных платформах
- настройка таргетированной рекламы
- настройка контекстной рекламы
- SEO-маркетинг для блога будет вестись параллельно с контент-маркетингом группы в ВКОНТАКТЕ.

ЗАКЛЮЧЕНИЕ ПО РАЗДЕЛУ СТАРТАП-КОНЦЕПЦИЯ

Платформа ПАРУС на базе алгоритма выбора образования представляет собой инновационный и перспективный проект, который может превратиться в прибыльный бизнес. Помимо этого, платформа поможет улучшить жизнь людей, помогая им с выбором образования. Не просто образованием, а образованием в согласии с их интересами.

Также платформа имеет прекрасные возможности для дополнительной монетизации, а именно: продажа рекламы, рейтинговая система онлайн-школ, помощь в подборе персонала.

Примерный срок окупаемости проекта составит порядка 25 месяцев с начала разработки. При этом этапы разработки и запуска платформы были учтены так, чтобы компания не ушла в «минус».

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа		ФИО	
0BM12		Гузеев Евгений Викторович	
Школа	ИЯТШ	Отделение (НОЦ)	Отделение экспериментальной физики
Уровень образования	магистратура	Направление/специальность	01.04.02 Прикладная математика и информатика

Тема ВКР:

Алгоритм и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения	
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	<i>Объект исследования:</i> разработка алгоритма <i>Область применения:</i> принятие решения студентом <i>Рабочая зона:</i> офис <i>Размеры помещения климатическая зона</i> * 15 м x 5 м x 4 м <i>Количество и наименование оборудования рабочей зоны:</i> компьютеры <i>Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне</i> разработка алгоритма на компьютере
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.	- Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018) - ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования.
2. Производственная безопасность при разработке проектного решения: - Анализ выявленных вредных и опасных производственных факторов - Расчет уровня опасного или вредного производственного фактора	Вредные факторы: - Отклонение показателей микроклимата; - Отсутствие или недостатки необходимого искусственного освещения Опасные факторы: - Повышенный уровень электромагнитных излучений; - Опасные и вредные производственные факторы, связанные с электрическим током

3. Экологическая безопасность <u>при разработке проектного решения</u>	Воздействие на литосферу: образования отходов при написании работы Воздействие на гидросферу: энерго и теплопотребление _____ Воздействие на атмосферу: энерго и теплопотребление _____
4. Безопасность в чрезвычайных ситуациях <u>при разработке проектного решения/</u>	Возможные ЧС аварии, пожары Наиболее типичная ЧС пожар
Дата выдачи задания для раздела по линейному графику	
01.03.2023	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Сечин Андрей Александрович	к.т.н		01.03.2023

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0BM12	Гузеев Евгений Викторович		01.03.2023

4 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

Социальная ответственность и охрана труда являются важнейшими составляющими любой деятельности, в особенности производственной, т.к. непосредственно связаны со здоровьем и жизнью человека.

Социальная ответственность - это сознательное отношение субъекта социальной деятельности к требованиям социальной необходимости, гражданского долга, социальных задач и, норм и ценностей, понимание осуществляемой деятельности для определенных социальных групп и личностей, для социального прогресса общества.

Актуальностью исследования «Алгоритм и его программная реализация для исследования принятия решения абитуриентами выбора направления обучения» является то, что специальность или направление обучения студента коррелирует с его наклонностями. А значит, что студент будет принимать решения по выбору специальности или направления образования более эффективно.

4.1 Правовые и организационные вопросы обеспечения безопасности

4.1.1. Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства

Согласно ТК РФ, N 197 -ФЗ работник аудитории 220, 11 корпуса ТПУ имеет право на:

- рабочее место, соответствующее требованиям охраны труда;
- обязательное социальное страхование от несчастных случаев на производстве и профессиональных заболеваний в соответствии с федеральным законом;
- отказ от выполнения работ в случае возникновения опасности для его жизни и здоровья вследствие нарушения требований охраны труда, за исключением случаев, предусмотренных федеральными законами, до устранения такой опасности;
- обеспечение средствами индивидуальной и

коллективной защиты в соответствии с требованиями охраны труда за счет средств работодателя;

- внеочередной медицинский осмотр в соответствии с медицинскими рекомендациями с сохранением за ним места работы (должности) и среднего заработка во время прохождения указанного медицинского осмотра;

Рабочее место в аудитории 220, 1 корпуса ТПУ должно соответствовать требованиям ГОСТ 12.2.032 - 78. Оно должно занимать площадь не менее 4,5 м², высота помещения должна быть не менее 4 м, а объем - не менее 20 м³ на одного человека. Высота над уровнем пола рабочей поверхности, за которой работает оператор, должна составлять 720 мм. Оптимальные размеры поверхности стола 1600 х 1000 кв. мм. Под столом должно иметься пространство для ног с размерами по глубине 650 мм. Рабочий стол должен также иметь подставку для ног, расположенную под углом 15° к поверхности стола. Длина подставки 400 мм, ширина - 350 мм. Удаленность клавиатуры от края стола должна быть не более 300 мм, что обеспечит удобную опору для предплечий. Расстояние между глазами оператора и экраном видеодисплея должно составлять 40 - 80 см. Рабочее место должно быть скомпоновано так, чтобы все операции работника выполнялись в пределах зоны досягаемости моторного поля в вертикальной плоскости (Рисунок 4.1) и в горизонтальной плоскости (Рисунок 4.2).

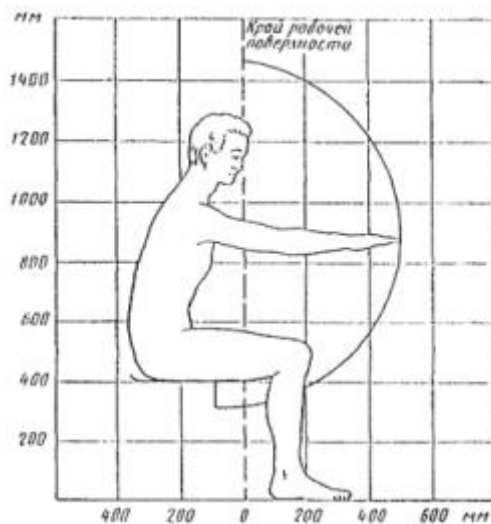


Рисунок 4.1 – Зона досягаемости моторного поля в вертикальной плоскости

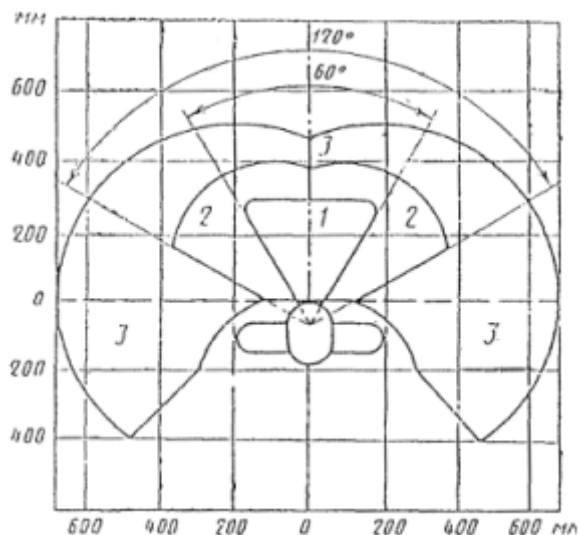


Рисунок 4.2 – Зона досягаемости моторного поля в горизонтальной плоскости.

1 - зона для размещения наиболее часто используемых предметов

2 - зона для размещения часто используемых предметов

3 - зона для размещения редко используемых предметов

Рабочее место сотрудника аудитории 220, 11 корпуса ТПУ соответствует требованиям ГОСТ 12.2.032 - 78.

4.2. Производственная безопасность

В учебной аудитории, имеются опасные и вредные производственные факторы. В таблице 4.1 приведены возможные опасные и вредные производственные факторы, влияющие на человека при исследовании.

4.2.1. Производственная безопасность

В таблице 4.1 приведены возможные опасные и вредные производственные факторы.

Таблица 4.1 – Опасные и вредные производственные факторы

Факторы	Нормативные документы
1.Отклонение показателей микроклимата	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [40]
2.Недостаточная освещенность рабочей зоны	СП 52.13330.2016 Естественное и искусственное освещение. [41]
3. Повышенный уровень электромагнитных излучений	ГОСТ 12.1.006 - 84 ССБТ. Электромагнитные поля радиочастот. Общие требования безопасности. [42]
4. Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	ГОСТ 12.1.038-82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов. [43]

Отклонение параметров микроклимата

Источник возникновения фактора – энергозатраты организма 151 – 200 (ккал/ч) [35], связанные с постоянной ходьбой, перемещением мелких (до 1 кг) изделий или предметов в положении стоя или сидя и требующие определенного физического напряжения, при работе в аудитории ТПУ.

При неблагоприятном микроклимате нарушается сердечный ритм, снижается концентрация, колеблется температура тела.

Поддержание оптимальных величин температуры и влажности воздуха рабочей зоны обеспечивается приточной и вытяжной вентиляцией, которая рассчитана на поглощение тепловыделений. Вентиляторы включаются автоматически при достижении температуры в помещении 33°C и отключаются при температуре 25 °С.

Допустимые параметры микроклимата приведены в таблице 4.2.

Таблица 4.2 Допустимые параметры микроклимата аудитории

Период года	Категория работ	Температура воздуха, °С		Температура поверхностей, t°С	Относительная влажность воздуха, φ%	Скорость движения воздуха, м/с
		Диапазон ниже оптимальных величин t° _{опт}	Диапазон выше оптимальных величин t° _{опт}			
Холодный	Iб	17,0 - 18,9	21,1 - 23,0	20 - 24,0	40 - 60	0,1
Теплый	Iб	16,0 - 18,9	22 - 24,0	21 - 25,0	40 - 60	0,1

В аудитории проводится ежедневная влажная уборка и систематическое проветривание после каждого часа работы на ЭВМ.

Недостаточная освещенность рабочей зоны

Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300 - 500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк .

В качестве источников света применяются светодиодные светильники или металлогалогенные лампы (используются в качестве местного освещения).

Повышенный уровень электромагнитных излучений

Источником повышенного уровня электромагнитных излучений являются мониторы наблюдения за технологическим процессом.

При длительном воздействии ЭМИ различных диапазонов длин волн при умеренной интенсивности (выше ПДУ) могут появиться головные боли, повышение или понижение давления, урежение пульса, изменение проводимости в сердечной мышце, нервно-психические расстройства, быстрое развитие утомления.

Нормирование ЭМИ радиочастотного диапазона проводится по ГОСТ 12.1.006 - 84 для производственной среды. Для защиты от влияния ЭМИ рекомендуется отдых и защитные очки.

Поражение электрическим током

Для предотвращения поражения электрическим током, где размещаются рабочее место с ЭВМ в аудитории 220, 11 корпуса ТПУ, оборудование оснащено защитным заземлением, занулением в соответствии с техническими требованиями по эксплуатации. Напряжение для питания ЭВМ 220 В, для серверного оборудования 380 В. По опасности поражения электрическим током помещение 220, 11 корпуса ТПУ относится к первому классу – помещения без повышенной опасности.

Основными непосредственными причинами электротравматизма, являются: 1) прикосновение к токоведущим частям электроустановки, находящейся под напряжением в случае пробоя изоляции; 2) прикосновение к металлическим конструкциям электроустановок, находящимся под напряжением; 3) ошибочное включение электроустановки или несогласованных действий обслуживающего персонала; 4) поражение шаговым напряжением.

Основными техническими средствами защиты, согласно ПУЭ [39], являются защитное заземление, автоматическое отключение питания, устройства защитного отключения, изолирующие электрозащитные средства, знаки и плакаты безопасности.

1. В качестве заземляющего устройства (вертикальные электроды) используем стальные трубы диаметром $d = 55$ мм, в качестве соединяющего элемента – стальная полоса шириной $b = 50$ мм.

2. Сопротивлению грунта в районе размещения установки или устройства.

Таблица 4.3 - Исходные данные для расчета

Вид заземления	контурное
Длина заземлителя l , м	2,7
Глубина заземлителя в грунте h , м	0,65
Сезонный коэффициент K_c	2,0
Удельное сопротивление земли ρ , Ом·м	70
Диаметр d , мм	55
Ширина соединительной полосы b , мм	50
Допустимое сопротивление системы заземления по ПУЭ $R_{з.у.}$, Ом	4
Уровень напряжения, В	220-380
Коэффициент экранирования	0,58

3. Величина электрического сопротивления растекания тока в грунт с одиночного заземлителя:

$$R_3 = 0,366 \cdot \frac{\rho \cdot K_c}{l} \left(\lg \cdot \frac{2 \cdot l}{d} + 0,5 \lg \cdot \frac{4 \cdot t + l}{4 \cdot t - l} \right) =$$

$$0,366 \cdot \frac{70 \cdot 2}{2,7} \left(\lg \cdot \frac{2 \cdot 2,7}{0,055} + 0,5 \lg \cdot \frac{4 \cdot 2 + 2,7}{4 \cdot 2 - 2,7} \right) = 38,51 \text{ Ом},$$

где,

$\rho = 70$ Ом-удельное сопротивление грунта,

$K_c = 2$ - коэффициент сезонности,

$l = 2,7$ м – длина заземлителя,

$d = 0,055$ м – диаметр заземлителя

$t = h + 0,5l = 0,65 + 0,5 \cdot 2,7 = 2$ м

4. Число заземлителей без взаимных помех, получаемых друг от друга, без так называемого явления «экранирования»:

$$n' = \frac{R_3}{R_{3,y}} = \frac{38,51}{4} = 9,62 \approx 10.$$

5. Число заземлителей с коэффициентом экранирования:

$$n = \frac{n'}{\eta_3} = \frac{10}{0,58} = 17,24 \approx 18.$$

Принимаем расстояние между заземлителями $a = l = 2,7$ м.

6. Длина соединительной полосы:

$$l_n = 1,05 \cdot n \cdot a = 1,05 \cdot 18 \cdot 2,7 = 51 \text{ м}.$$

7. Значение сопротивления растекания тока с соединительной полосы:

$$R_3 = 0,366 \cdot \frac{\rho \cdot K_c}{l} \left(\lg \cdot \frac{2 \cdot l_n^2}{b \cdot h} \right) = 0,366 \cdot \frac{70 \cdot 2}{51} \left(\lg \cdot \frac{2 \cdot 51^2}{0,05 \cdot 0,65} \right) = 5,1 \text{ Ом}.$$

8. Полное сопротивление системы заземления:

$$R_{3y} = \frac{R_3 \cdot R_{\Pi}}{R_3 \cdot \eta_{\Pi} + R_3 \cdot \eta_3 \cdot n} = \frac{38,51 \cdot 5,1}{38,51 \cdot 0,51 + 5,1 \cdot 0,58 \cdot 18} = 2,63 \text{ Ом},$$

где,

$\eta_{\Pi} = 0,51$ - коэффициент экранирования полосы.

Таким образом, сопротивление $R_{\Sigma} = 2,63 \text{ Ом}$ не превышает 4 Ом . Следовательно, диаметр заземлителя $d = 55 \text{ мм}$ при числе заземлителей $n = 18$ является достаточным для обеспечения защиты при контурной схеме расположения заземлителей.

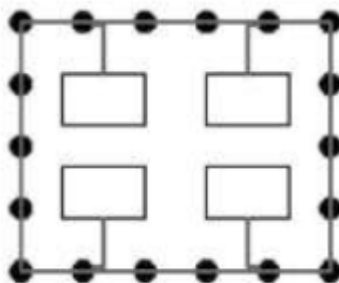


Рисунок 7.1 – Схема полученного контурного заземления

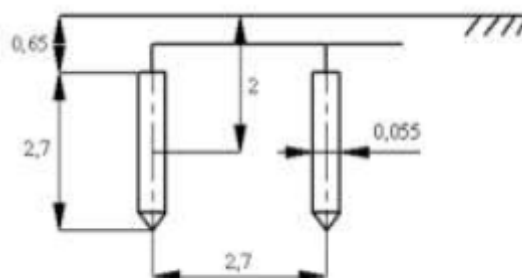


Рисунок 7.2 – Схема расположения заземлителей

Разработанные мероприятия и расчеты обеспечивают безопасную эксплуатацию электроустановок в аудитории 41, 4 корпуса ТПУ.

4.3. Экологическая безопасность

Влияние на атмосферу и гидросферу

При использовании компьютера для разработки потребляется электроэнергия. Выработка электроэнергии осуществляется на ТЭЦ и ГРЭС. При выработке электроэнергии за счет сжигания топлива в воздух поступают различные загрязнения.

Для уменьшения выбросов на ТЭС применяют различные фильтры и разные технологии сжигания.

На гидросферу также влияет по большей части выработка тепла и электроэнергии. Гидросфера загрязняется обмывочными водами котлов, замасляемыми водами, сброс тепловых потоков.

Чтобы минимизировать сброс обмывочных и замасленных вод в гидросферу, чаще всего используют очистку и масляные ловушки.

Влияние на литосферу

Процесс исследования представляет из себя работу с информацией, такой как технологическая литература, статьи, ГОСТы и нормативнотехническая документация, а также разработка математической модели с помощью различных программных комплексов. Таким образом процесс исследования имеет влияние негативных факторов на окружающую среду. Таких как отходы – использованная бумага, использованные шариковые ручки.

Использованная бумага и пластиковые шариковые ручки утилизируется, как вторичное сырье – изготовление картона, пластиковой тары и.т.д. Процесс утилизации негативно влияет на атмосферу, выделяя в нее углекислый газ.

Существующая система фильтрации не может на 100% избавить от выделения вредных веществ в атмосферу. Что касается методов по защите литосферы, то используются следующие методы:

- Энергосбережение;
- Сортировка мусора.

4.4 Безопасность в чрезвычайных ситуациях

4.4.1. Анализ вероятных ЧС, которые может инициировать объект исследований и обоснование мероприятий по предотвращению ЧС

При проектировании и эксплуатации программного продукта возможны следующие чрезвычайные ситуации в рабочей зоне:

- техногенные - производственные аварии, пожары;

Наиболее типичным и вероятным видом чрезвычайной ситуации является пожар в рабочей зоне. Наиболее вероятные причины пожара связаны с неисправностью или ненадлежащей эксплуатации электроприборов – короткое замыкание или перегрузки по току, нарушение работником правил эксплуатации. В соответствии с СП 12.13130.2009 помещение лаборатории относится к категории В1 зона П-Па. Перед тушением необходимо отключить общее электроснабжение. Для тушения можно использовать огнетушители разного типа, но, если отключить электроснабжение не удастся, необходимо применять только порошковые и углекислотные. Требования к эксплуатации огнетушителей приведены в СП 9.13130.2009.

4.4.2. Меры по предупреждению возникновения пожара

При проведении исследований наиболее вероятной ЧС является возникновение пожара в помещении 220, 11 корпуса ТПУ. Пожарная безопасность должна обеспечиваться системами предотвращения пожара и противопожарной защиты, в том числе организационно-техническими мероприятиями. Основные источники возникновения пожара:

- 1) Неработоспособное электрооборудование, неисправности в проводке, розетках и выключателях. Для исключения возникновения пожара по этим причинам необходимо вовремя выявлять и устранять неполадки, а также проводить плановый осмотр электрооборудования.
- 2) Электрические приборы с дефектами. Профилактика пожара включает в себя своевременный и качественный ремонт электроприборов.
- 3) Перегрузка в электроэнергетической системе (ЭЭС) и короткое замыкание в электроустановке.

Под пожарной профилактикой понимается обучение пожарной технике безопасности и комплекс мероприятий, направленных на предупреждение пожаров.

Пожарная безопасность обеспечивается комплексом мероприятий:

- обучение, в т.ч. распространение знаний о пожаробезопасном поведении (о необходимости установки домашних индикаторов задымленности и хранения зажигалок и спичек в местах, недоступных детям);
- пожарный надзор, предусматривающий разработку государственных норм пожарной безопасности и строительных норм, а также проверку их выполнения;
- обеспечение оборудованием и технические разработки (установка переносных огнетушителей и изготовление зажигалок безопасного пользования).

В соответствии с ТР «О требованиях пожарной безопасности» для административного жилого здания требуется устройство внутреннего противопожарного водопровода.

Согласно ФЗ-123, НПБ 104 - 03 «Проектирование систем оповещения людей о пожаре в зданиях и сооружениях» для оповещения о возникновении пожара в каждом помещении должны быть установлены дымовые оптикоэлектронные автономные пожарные извещатели, а оповещение о пожаре должно осуществляться подачей звуковых и световых сигналов во все помещения с постоянным или временным пребыванием людей.

Аудитория 41, 4 корпуса ТПУ оснащена первичными средствами пожаротушения: огнетушителями ОУ-3 1шт., ОП-3, 1шт. (предназначены для тушения любых материалов, предметов и веществ, применяется для тушения ПК и оргтехники, класс пожаров А, Е.).

Согласно НПБ 105-03 помещение, предназначенное для проектирования и использования результатов проекта, относится к типу П - 2а.

4.4.3 Действия в случае возникновения пожара

В случае возникновения пожара необходимо:

- спокойно оценить ситуацию и принять срочные меры по предотвращению распространения огня;

- вызвать пожарных, сообщив точный адрес места возгорания и ФИО, вызывающего;
- отключить общее электроснабжение;
- попытаться потушить пожар при помощи первичных средств пожаротушения;
- если самостоятельно справиться с огнем не удаётся, следует включить сигнал пожарной тревоги и, согласно плану, приступить к эвакуации;
- встретить прибывшую пожарную команду и обеспечить для неё беспрепятственный доступ и пути подъезда к месту пожара.

Вывод по разделу

В результате выполнения задания раздела «Социальная ответственность и ресурсосбережение» ВКР были выявлены и проанализированы вредные факторы при разработке алгоритма.

Были установлены правовые и организационные вопросы обеспечения безопасности, характерные для рабочей зоны.

Проработаны организационные мероприятия при компоновке рабочей зоны.

Разработаны мероприятия по снижению воздействия вредных и опасных факторов.

Был рассмотрен характер воздействия исследуемого решения на окружающую среду. Были выявлены предполагаемые источники загрязнения окружающей среды, возникшие в результате реализации предлагаемых в магистерской диссертации решений.

Можно сделать вывод, что создание магистерская диссертация не является экологически безвредным действием. Так как процесс сопровождается созданием отходов от проектной деятельности.

А также процесс написания магистерской диссертации не является абсолютно безвредным для человека. Так как в процессе написания магистерской диссертации человека сопровождают такие вредные факторы, как электромагнитное излучение от ПК и плохая освещенность рабочей зоны.

Даны общие рекомендации по безопасности, соблюдая которые можно, не только сохранить здоровье, но и увеличить эффективность.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы было выявлено:

- 1) Пользователь состоит в группах ВКОНТАКТЕ, связанных с его профессиональными интересами.
- 2) Некоторые группы не относятся к интересам, но такие группы обобщены на многие специальности. К таким группам можно отнести: юмор, новости, хобби, путешествия.
- 3) При разработке первого алгоритма создавался классификатор баллов согласно размеченным данным. Наилучшего результата по метрике F1 достиг алгоритм градиентного бустинга Catboost. Значение метрики составило 0.771. F1 у алгоритма логистической регрессии и ближайших соседей составило 0,729 и 0,726 соответственно. Использование предобученных моделей в модели-трансформера показало результат хуже этих алгоритмов. Это связано с подбором предобученных моделей.
- 4) На классификации направления обучения алгоритм 1 показал результат по метрике F1 0.63.
- 5) Из алгоритма 1 как следствие получился алгоритм 2. Алгоритм представлял собой функцию сбора и препроцессинга данных, стек обученных моделей. Результат работы алгоритма по метрике F1 составил более 0.8
- 6) На основе алгоритма разработана концепция платформы, которая работает следующим образом. Пользователь, ищущий образование заходит на сайт с платформой вводит свой ID и на выходе получает, результаты той образовательной траектории, которое ближе к его интересам.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. А. Фещенко. Как помочь подросткам и университетам найти друг друга?., [Электронный ресурс] URL: <https://urbanru.ru/feshhenko-kak-pomoch-podrostkam-universitetam-najti-drug-druga/> Дата обращения (28.03.2023)
2. Pavel Kiselev., Career guidance based on machine learning: social networks in professional identity construction., *Procedia Computer Science* 169 (2020) 158–163
3. Смирнов И.Б., Сивак Е.В., Козьмина Я.Я. В поисках утерянных профилей: достоверность данных ВКонтакте и их значение в образовательных исследованиях // *Вопросы образования*. 2016. № 4. С. 106–122.
4. Фещенко А.В., Гойко В.Л., Степаненко А.А., Суханова Е.А., Мацута В.В., Киселев П.Б. Методы и инструменты выявления перспективных абитуриентов в социальных сетях // *Открытое и дистанционное образование*. 2017. № 4 (68). С. 45-52.
5. В.В. Кашпур, Е.Ю. Петров, В.Л. Гойко, А.В. Фещенко., Возможности использования цифровых следов для образовательных достижений студентов., *Вестник Томского государственного университета., Философия. Социология. Политология*. 2021 г.
6. Fail M. Gafarov, Konstantin S. Nikolaev, Pavel N. Ustin 1, Andrey A. Berdnikov, A Complex Neural Network Model for Predicting a Personal Success based on their Activity in Social Networks., *EURASIA Journal of Mathematics, Science and Technology Education*, 2021, 17(10), em2010
7. Груздев А.В. Предварительная подготовка данных в Python: Том 1. Инструменты и валидация. –М.: ДМК Пресс, 2023. –816 с.: ил.
8. Хейдт М., Груздев А. В. Изучаем pandas / пер. с англ. А. В. Груздева. – М.: ДМК Пресс, 2019. – 682 с.: ил.
9. Груздев А., Курс «Машинное обучение» в R, Python и H₂O 5.6»., Сборник переведенных статей., 391 с., 2019 г.
10. Векторизация текста с использованием методов интеллектуального анализа данных., Ю.А Кравченко., А.М Мансур., Ж.Х. Мохаммед., *Известия ЮФУ. Технические науки*, 2019 г
11. Груздев А.В. Предварительная подготовка данных в Python: Том 2. План, примеры и метрики качества.. –М.: ДМК Пресс, 2023. –816 с.: ил.
12. Саваш Йылдырым, Мейсам Асгари-Ченаглу., Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов

обработки естественного языка / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2022. – 318 с.: ил

13. Topic Modeling in Python: Latent Dirichlet Allocation (LDA) [Электронный ресурс]., URL: <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0> ., Дата обращения (11.05.2023)

14. Мэрфи К.П Вероятностное машинное обучение: введение / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 940 с.: ил.

15. Бурков А. Инженерия машинного обучения / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 306 с.: ил.

16. Монарх (Манро) Р.Х20 Машинное обучение с участием человека / пер. с англ. В. И. Бахура. – М.: ДМК Пресс, 2022. – 498 с.: ил.

17. Джоши Прадик Искусственный интеллект с примерами на Python. : Пер. с англ. - СПб. : ООО "Диалектика", 2019. - 448 с. - Парал. тит. англ.

18. Лимановская, О. В. Основы машинного обучения : учебное пособие / О.В. Лимановская, Т. И. Алферьева ; Мин-во науки и высш. образования РФ. — Екатеринбург : Изд-во Урал. ун-та, 2020. — 88 с.

19. Алгоритмы Data Science и их практическая реализация на Python : учебное пособие / А. В. Протоdjяконов, П. А. Пылов, В. Е. Садовников. - Москва ; Вологда : Инфра-Инженерия, 2022. - 392 с.

20. Итоги четвёртого квартала 2022 года ВКонтакте [Электронный ресурс]. URL: <https://vk.com/press/q4-2022-results> (Дата обращения 20.04.2023)

21. Пользователи соцсетей в России: статистика и портреты аудитории [Электронный ресурс]. URL: https://vk.com/@cerebro_vk-polzovateli-socsetei-v-rossii-statistika-i-portrety-auditori (Дата обращения 10.05.2023)

22. Федеральная служба государственной статистики. Численность населения Российской Федерации по полу и возрасту на 1 января 2020 года. [Электронный ресурс]. URL: https://gks.ru/bgd/regl/b20_111/Main.htm (Дата обращения 10.05.2023)

23. Объем рынка онлайн-образования России (2021-2022) PAM-TAM-SAM-SOM. Алексей Ручков. [Электронный ресурс]. URL:

- https://vc.ru/marketing/458127-obem-rynka-onlayn-obrazovaniya-rossii-2021-2022-pam-tam-sam-som#_ftn1 (Дата обращения 10.05.2023)
24. Программа лояльности SKILLBOX [Электронный ресурс]. URL: <https://partners.skillbox.ru/referral> (Дата обращения 10.05.2023)
25. Исследование рынка онлайн-образования. Нетология. [Электронный ресурс] URL: https://netology.ru/edtech_research_2022 (Дата обращения 20.04.2023)
26. Федеральный институт промышленной собственности. Программы для ЭВМ, БД. [Электронный ресурс] URL: <https://fips.ru/to-applicants/software-and-databases/> (Дата обращения 10.05.2023)
27. Edtech-рынок закончил год ростом почти в 18%. РБК ТRENДЫ. [Электронный ресурс] URL: <https://trends.rbc.ru/trends/education/63f46feb9a794744929f5d29>
28. EDUCATION TECHNOLOGY: A GIANT AWAKENS. [Электронный ресурс] URL: <https://rizeetf.com/2021/03/education-technology-a-giant-awakens/> (Дата обращения 12.05.2023)
29. Knewton. [Электронный ресурс]. URL: <https://www.knewton.com/> (Дата обращения 10.05.2023)
30. CollegeVine. [Электронный ресурс]. URL: <https://www.collegevine.com> (Дата обращения 10.05.2023)
31. Cialfo. [Электронный ресурс]. URL: <https://www.cialfo.co/> (Дата обращения 10.05.2023)
32. Edurank. [Электронный ресурс]. URL: <https://edurank.org/> (Дата обращения 10.05.2023)
33. Построение бизнес-моделей: Настольная книга стратега и новатора / Александр Остервальдер, Ив Пинье; Пер. с англ.—2-е изд. М: Альпина Паблишер, 2023.—288 с.
34. Тестирование бизнес-идей. Остервальдер Александр, Блэнд Дэвид. М: Альпина Паблишер, 2020.—354 с.

35. Стартап. Настольная книга основателя/ Бланк Стив, Дорф Боб. М: Альпина Паблишер, 2021.–616 с.

36. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. От 27.12.2018)

37. Федеральный закон от 29.11.2010 № 326-ФЗ (ред. От 24.02.2021) «Об обязательном медицинском страховании в Российской Федерации»

38. ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования.

39. ГОСТ 12.0.003-2015 ССБТ. Опасные и вредные производственные факторы. Классификация.

40. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания

41. СП 52.13330.2016 Естественное и искусственное освещение.

42. ГОСТ 12.1.006-84 ССБТ. Электромагнитные поля радиочастот. Общие требования безопасности.

43. ГОСТ 12.1.038-82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов.

44. ПУЭ: правила устройства электроустановок. Издание 7.

45. СП 12.13130.2009 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности

СПИСОК ПУБЛИКАЦИЙ

1. Мультиклассовая классификация корпуса смешанных текстов алгоритмами машинного обучения., Е.В. Гузеев, М.Е. Семенов., Материалы XX Международная конференция «Перспективы развития фундаментальных наук» принято в печать., 25-28 апреля 2023г., г.Томск

ПРИЛОЖЕНИЕ А
(справочное)

**Algorithm and its software implementation to study the decision-making of
applicants to choose the direction of education**

Студент

Группа	ФИО	Подпись	Дата
ОВМ12	Гузеев Е.В		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Семенов М.Е	к. ф.-м. н.		

Консультант-лингвист отделения иностранных языков ШБИП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель	Смирнова У.А			

INTRODUCTION

Social media plays an important role in people's lives nowadays. People learn news from social networks, they are shown various advertisements, many people form clubs according to their interests, date of birth, age very many interactions between derive from the information received. In social networks based on the information collected recommendation systems are based. They recommend certain content to users or recommend a user of a certain person.

Social media is used everywhere in modern life. Businesses, for example, look for their target audience from social media, which can optimize the cost of their advertising budget. Some media outlets can spread certain news with lightning speed, using the viral component of news content. Social media can identify destructive communities or communities that pose a threat in today's world. Since people use social media almost daily and leave, certain social traces interests, reposts, music preferences, subscriptions. Based on this data, a rough digital portrait of a person can be obtained.

Such a portrait can be used quite extensively from marketing interactions to identifying the social behavior of users.

Now in the educational policy of the Russian Federation there is a tendency to identify more talented students, their professional orientation and prediction of their professional success. Traditional data attributes for solving these tasks are data generated by users in e-learning systems, as well as psychological and sociological surveys. However, these data are not enough to reveal a more complete picture of the student's world and make educational predictive models. Therefore, social networks can help here. A user of a social network, if he leads an active life (social circle, circle of interests, musical preferences) can be said to show his/her preferences and interests to a sufficient extent. And these preferences and interests can be used as a feature space for making predictive models.

Such predictive models based on a user's interests in a social network can help with the choice of major or field of study. At the same time, if a person studies and works in accordance with their interests, education will go easier and much more effectively for them.

The aim of this master's thesis is to develop an algorithm and its software implementation to study applicants' decision-making in choosing a field of study, based on the methods of social network analysis and machine learning.

Scientific and practical novelty and significance of the work:

The proposed method solves the problem of choosing the direction of study by applicants, students and professionals. Based on the developed algorithm will build a digital platform that will recommend users the most relevant option for learning, depending on the interests of the user of the social network VKONTAKTE. To achieve this goal it is necessary to work out the following tasks:

1. Provide a literature review on social network analysis and the use of data as markers for identifying professional interests.
2. Consider methods of multi-class classification of texts

1 REVIEW OF SCIENTIFIC AND TECHNICAL LITERATURE ON THE STUDY

1.1 Digital footprints in educational activities

Studying a student's digital footprint can improve learning analytics and prediction in several ways:

First, the digital footprint offers a wealth of open-ended user data on personality traits such as cognitive, motivational, and psychological traits. This data is generated naturally by the student rather than artificially, such as through tests or surveys.

Second, analyzing a student's digital footprint can help assess his or her educational potential at the beginning of university. Unlike existing models that rely on current academic performance, digital footprint analysis can predict the success of students who have just entered university and have yet to take their first session.

Third, digital footprints provide a new way to assess and analyze educational achievement that is not formally captured within the university's curriculum. As a result, there is a growing need for analytical tools and models that can work effectively with digital footprints to improve the quality of management decisions in the educational process and students' educational trajectories.

In 2019, the authors of Tomsk State University [1] conducted research on the creation of an algorithm for finding suitable high school students in VKontakte. According to the results, the authors' algorithm determines the level of creativity, intelligence and motivation.

An example of the algorithm is shown in Figure 1.



Fig 1. Example of an algorithm to find high school students for a university [1]

The authors [1] investigated both the text content on user walls and their subscriptions. In general, they investigated about 9 thousand educational communities out of 1 million entertainment communities. The results of the study are shown in Figure 2.

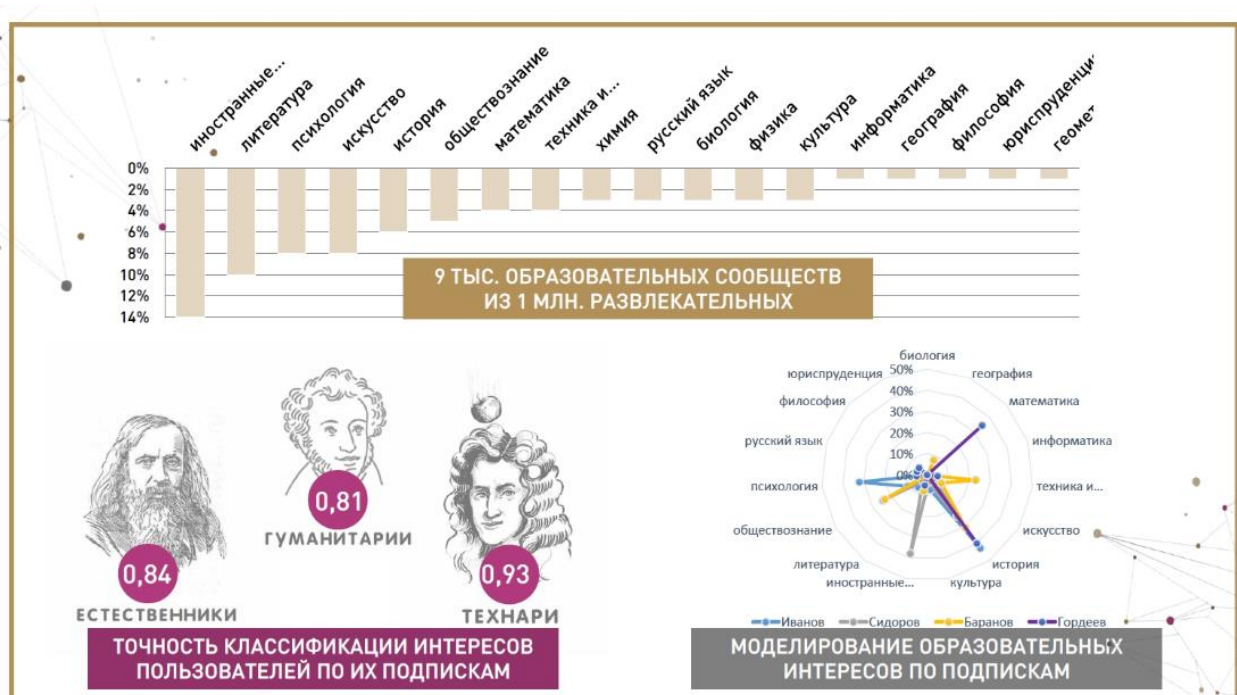


Fig 2. Results of user subscription analysis [1]

They also used network analysis in their work to identify levels of difference in creativity. The results are presented in Figure 3.



Fig 3. Results of network analysis [1]

Based on the network analysis, the authors concluded that networks of high school students with low creativity usually consist of 120-220 friends, these networks are exceptionally segmented, each graph consists of an average of 13-25 clusters. From them it is possible to distinguish 5-6 large communities, which are practically not connected among themselves; it is also evidenced by high coefficient of modularity at low density of a network [1].

Schoolchildren with a high level of creativity tend to form a large number of contacts, their network consists on average of 260-400 friends, these networks have a pronounced segmentary character. However, different clusters have a significant number of connections with each other.

In the source [2] the authors conduct a study of professional interests of a social network user on the basis of their data, supplemented by professional surveys to identify interests. Based on the results of the study, the authors made a predictive model. The ROC-AUC metric of this model varied from 0.7 for the management sphere to 0.85 for the production sphere. In addition, the authors [2] show that artificial intelligence cannot help with career building, but can help in the awareness of professional interests.

Since about 2015, the growth of publications on the topic of analyzing digital footprints in educational activities begins. For example, in the source [3], the author

shows the relationship between a student's presence in the social network "VKontakte" and his or her grade point average.

In the article [4], the authors consider the methodology of identifying promising applicants in the social network. According to the study, the authors identified three areas for finding promising applicants in the following areas: humanities, exact sciences, natural sciences. The authors divided all the research into four directions [4]:

In the first direction, the authors made a hypothesis about the relationship of schoolchildren's interests to the content in the social network "VKONTAKTE". In the second area, the authors tested the hypothesis that the personal pages of users can reflect the interests of applicants. In the third direction, the potential of social networks for identifying important psychological qualities of an applicant was studied. The fourth direction assumed the presence of features in the structure of social connections in gifted high school students [4].

The results of their study show, the method of predicting educational interests and traits of giftedness by user subscriptions gave the best results. It can also be deduced from the conclusions that the giftedness of applicants in the social network is quite difficult to catch because of the uneven contexts of user subscriptions.

In the source [5], the authors combined the use of digital traces, learners with their traces in the learning environment LMS Moodle. The data obtained from the social network and the authors used educational environment as the feature space of the model to predict the educational achievements of students. To build a model, they used modern libraries based on the algorithm of gradient binning of the following libraries: CatBoost and LightGBM. It is also worth noting the approach of the authors, in which they considered high educational interests. The authors proposed that students whose average score was above the boundary of the 65th percentile and whose average score was below the boundary of the 35th percentile were considered to have high educational achievement.

Source [6] discusses the prediction of a user's personal success based on social network activity. The authors used the nearest neighbor algorithm to divide users

into 5 clusters based on success. In the course of the work, they augmented the input data including, number of friends, number of followers, number of entries on the wall.

In computational experiments [6] it was found that there was a positive effect on the accuracy of the neural network when more input parameters were added to train the neural network. In addition, it was confirmed that the distribution of the original sample of people into success clusters is correct, as there is a monotonic increase in the differences with increasing range between the two selected clusters. The separation of cluster 5 (the most professionally successful people) from all others did not show good results.

User characteristics vector

Each network user will have his own characteristics vector based on his interests. The vector of characteristics is a vector of a given dimension, which will characterize the user of the social network. For its construction, weighting coefficients, responsible for the contribution of each element of the available sets of information in the result, are used. The final values of these values, obtained by taking into account all available information about the user by summation with normalization, introduced to account such aspects as the possibility of providing different amounts of information by users, different levels of significance of some results relative to others and others, form a vector necessary for comparison. Each element of the vector corresponds to some interests of the user.

To exclude the influence of unconscious factors it is necessary to cluster users' interests and minimize the influence of interests that do not influence anything. Such interests include such topics as humor, games, auctions, ads, city news, etc. Since the social network "VKontakte" there is no clear clustering of groups, communities, but only a description of it. In this case, the task is clustering groups based on extraction of meaning from its description, that is, thematic modeling. This approach is quite time-consuming and its result should be tested.

It is worth noting that the feature space of the content of users is a textual features. Therefore it is expedient to consider models of classification on the basis of text data in the developed algorithm.

2 MULTI-CLASSIFICATION OF MIXED TEXTS CORPUS BY MACHINE LEARNING ALGORITHMS

Logistic regression when replacing the sigmoid function with a softmax function can be used for multiclass classification problems. The nearest-neighbor algorithm also handles these types of problems. To do this, it is necessary to find the k nearest neighbors of the data for input X and return the class to which most of the data belongs among k [7, 8]. If the assumption of feature independence is not violated, a polynomial naive Bayesian classifier can be used, which assumes that each class has its own set of parameters. The algorithm calculates the probability of occurrence of a class in the sample. If we consider algorithms based on decision trees, the random forest and gradient boosting algorithms are well suited. A random forest consists of several decision trees, each trained on a random subsample of data. Each tree is constructed by partitioning into two or more subgroups based on a particular attribute. The main idea of the gradient-based boosting is that each next decision tree takes into account the errors of the previous trees and tries to correct them. In this way, the quality of the predictions is gradually improved [9, 10].

2.1 transformer models for multiclass text classification problems

The effectiveness of transformational models in various nlp tasks, including text generation and classification, is of considerable interest. The attention mechanism is a crucial component of these models. However, even before transformers, the attention mechanism was used to improve conventional deep learning models such as RNN. To gain a full understanding of transformers and their importance in NLP, one must first study the attention mechanism [12].

This mechanism is based on the fact that models based on Rnn-based models, such as GRU or LSTM, have an information bottleneck For tasks such as neural machine translation (neural machine translation, NMT). These encoder-decoder-based models receive input as token-id. And subject it to recurrent processing (encoder). The processed intermediate representation is then fed into another recurrence unit (decoder) to extract the results [12].

To remove this bottleneck and match the decoder with the encoder, an attention mechanism was proposed that uses weights assigned to intermediate latent values. These weights determine the amount of attention the model should give to the input data at. At each decoding step. Before turning to mechanisms of attention based on weighted scalar products, it is necessary to understand what selfattention is. The self-attention mechanism schematically shown in Figure 10 is a basic form of the scaled self-attention mechanism (12).

This mechanism receives the input matrix X and performs evaluation of the attention of its elements. In our case X is a 3×4 matrix, where 3 is the number of tokens and 4 is the representation size. The matrix Q is called the query, K the key, and V the value. Multiplying X obtains these matrices and the matrices labeled Θ , Φ , and g , respectively. Multiplication of the query matrices (Q) and the key matrices (K) gives a matrix of attention weights. This mechanism can also be seen as a database, in which we use query and key to get a numerical estimate of the connections between different elements. Multiplying the attention weights and the matrix V gives the final result of this type of attention mechanism. The main reason it is called self-attention is its single input X ; the matrices Q , K , and V are computed from X .

Transformer models account for encoder annotations and hidden values from past layers. The transformer architecture has no recursive step-by-step flow; instead, it uses positional coding to have information about the position of each token in the input sequence. The layers in the first part of the encoder are fed with input data, which are combined representation values (initialized randomly) and fixed positional coding values; they propagate through the architecture, as shown in Figure 4 [12].

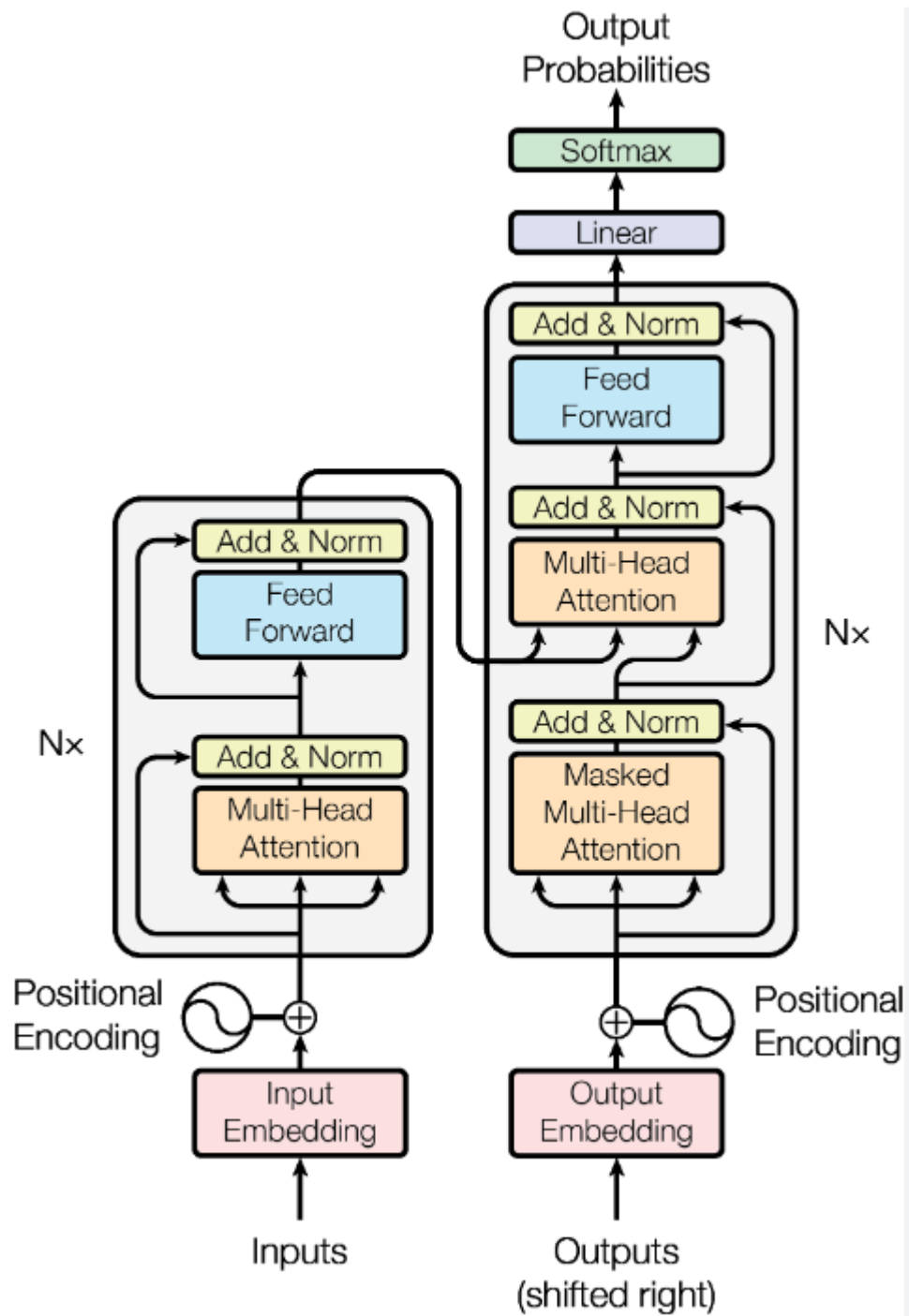


Fig 4. Generalized transformers model

The goal of this work is to conduct a comparative analysis of the quality of machine learning algorithms for the multiclass classification problem. To achieve the goal, we set and consistently solved the following tasks:

1. Collecting (parsing) raw data.
2. Data pre-processing.
3. Preparation of data for classification.
4. Choosing a quality metric for data classification.
5. Choosing the best algorithm based on a quality metric.

Experimental part. As input data, we used a dataset that we collected based on data from open groups in the social network VK. To test different algorithms, we formulated the following hypothesis "A user who is interested in applied computer science is partly a member of the social network communities of interest". This hypothesis was tested, based on the grouping and sorting of non-unique data, which shows that people are members of communities in one way or another related to programming or mathematics. A parser based on VK API was written to collect the data. For each user_id the data was collected for the 10 groups he is a member of. The resulting data table includes 7863 rows and 4 columns: description (group description), name (group name), activity (general topic of the group), group_id (group identifier). The original data had 54,5 % unique data, after deleting the duplicates there were 4365 rows. Next we partitioned the data into 4 classes: 0 - the group is unrelated, and 3 - the group is directly related to the target audience of interest. Class distribution (in fractions): 0 - 0,52; 1 - 0,37; 2- 0,073; 3 - 0,037. An example of the data after preprocessing is shown in Figure 5.

	description	activity	name	screen_name	target
0	Присылайте свои анекдоты в "Предложите новость...	Юмор	Азохнвей! Бесплатные щикарные анекдоты	azohen_vey	1
1	Познавательные мемы, широкий охват тем, уникал...	Юмор	Мемуары ценителей научных мемов	sciencemem	1
2	На Хабр Q&A можно получить ответ на вопрос по ...	Программирование	Хабр Q&A	toster_ru	4
3	Официальный канал сервиса РБК Инвестиции для т...	Финансы	РБК Инвестиции	rbc_investments	3
4	Помогаем новичкам стать веб-разработчиками, а ...	Программирование	Хекслет	hexlet	4

Fig. 5. Example of raw data.

Note that the sample is not balanced: the majority class is 14 times larger than the minority class, so we chose to evaluate the quality of classification - the metric F_1 , which is a harmonic mean of accuracy and completeness [7-9]:

$$F_1 = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

We converted the marked dataset into a bag of words, and then applied frequency analysis to the text. The following steps were performed to implement the bag of words:

1. Removed stop words in the text that do not carry any meaning.
2. Split the text data in the documents into words, reduced all the words to their original word form (lemmatization and tokenization process).
3. Formed a dictionary of all the words in the corpus of the document. We calculated the frequency of occurrence of each word.
4. We created a vector space from the text data in the form of a disjoint matrix.

The data were converted into a vector TFIDF space to form the feature set. This space can estimate the importance of a word within a particular document: a word is given a higher weight if it appears more often in a document, but is rarely found in the rest of the corpus. If a word appears frequently in a particular document, but rarely in the rest of the document, it is likely to describe the content of that document better [7]. The transformed vector space table has 4365 rows and 23464 columns.

	5789	5878	6098	6552	6686	7396	7629	7637	7761	7864	...	20614	20798	21046	21439	21804
0	0.070666	0.138949	0.539146	0.11321	0.074291	0.11321	0.132464	0.179032	0.138949	0.057835	...	0.093175	0.063489	0.11321	0.116779	0.072068
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.207997	...	0.000000	0.000000	0.000000	0.000000	0.194389

<4365x23464 sparse matrix of type '<class 'numpy.float64''>
with 115993 stored elements in Compressed Sparse Row format>

Fig 6. Example of data in vector TFIDF space.

Results. As a result, accuracy and completeness metrics were calculated for each algorithm and used to calculate the metric F_1 (Table 1). For each model, hyperparameters were selected to optimize the metric F_1 .

Table 1 – Results of the comparison of the quality of algorithms

Algorithm	Metric F_1
Logistic regression	0.729
Random Forest	0.709
Naive Bayes	0.706
Nearest neighbors	0.726
Gradient Boosting (CatBOOST)	0.771
Transformer model	0, 713

Conclusion. As a result of the study it was shown that, the algorithms of gradient binning, logistic regression and the nearest neighbor algorithm were the most successful in the task of multiclass classification. Due to the fact that the classification task is to find the probability of a relationship to a class, and hence the results can be improved by building an ensemble of models.

REFERENCE

1. A. Feshchenko. How to help teenagers and universities find each other?, [Electronic resource] URL: <https://urbanru.ru/feshhenko-kak-pomoch-podrostkam-universitetam-najti-drug-druga/> Access date (28.03.2023)
2. Pavel Kiselev, Career guidance based on machine learning: social networks in professional identity construction, *Procedia Computer Science* 169 (2020) 158-163
3. Smirnov I.B., Sivak E.V., Kozmina Ya.Y. In search of lost profiles: reliability of VKontakte data and their importance in educational research // *Voprosy Obrazovanie*. 2016. № 4. C. 106-122.
4. Feshchenko A.V., Goiko V.L., Stepanenko A.A., Sukhanova E.A., Matsuta V.V., Kiselev P.B. Methods and tools for identifying prospective applicants in social networks // *Open and Distance Education*. 2017. № 4 (68). C. 45-52.
5. V. V. Kashpur, E. Yu. Petrov, V.L. Goiko, A.V. Feshchenko, Possibilities of using digital traces for educational achievements of students, *Vestnik of Tomsk State University, Philosophy. Sociology. Political science*. 2021 г.
6. Fail M. Gafarov, Konstantin S. Nikolaev, Pavel N. Ustin 1, Andrey A. Berdnikov, A Complex Neural Network Model for Predicting a Personal Success based on their Activity in Social Networks, *EURASIA Journal of Mathematics, Science and Technology Education*, 2021, 17(10), em2010
7. Gruzdev A. V. *Data Prefabrication in Python: Volume 1. Tools and Validation*. -M.: DMK Press, 2023.- 816 c.
8. Burkov A. *Machine learning without extra words*. - SPb: Peter, 2020. - 192 c.
9. Murphy K.P. *Probabilistic machine learning: an introduction* / translated from English by A.A. Slinkin. -Moscow: DMK Press, 2022. - 940 c.
10. Burkov A. *Machine Learning Engineering* / translated from English by Slinkin A. A.. - Moscow: DMK Press, 2022. - 306 c.

11. Vasiliev Y. Natural Language Processing. Python and spaCy in practice. - SPb: Peter, 2021. - 256 c.
12. Savaş Yıldırım, Meysam Asgarı-Cenaglu, Mastering Transformer Architecture. Development of modern models using advanced natural language processing methods / translated from English. V. S. Yatsenkov. - Moscow: DMK Press, 2022. - 318 p.: ill.

ПРИЛОЖЕНИЕ Б ОСНОВНЫЕ ССЫЛКИ ПО РЕШЕНИЮ

1. Задачи парсинга данных на основе API вконтакте и задачи мультиклассовой классификации текстовых данных методами машинного обучения. URL:
<https://colab.research.google.com/drive/1wE17UaEl6qTyuqNwqsC1ZfAfZA8Tvmgo?usp=sharing>
2. Задачи мультиклассовой классификации текстовых данных на основе моделей-моделей-transformer. URL:
https://colab.research.google.com/drive/1Me_plOIONxJ6GfOHv-o-OorN8J4h03Bx?usp=sharing
3. Задача бинарной классификации с использованием текстовых признаков (Создание стека классификаторов) URL:
<https://colab.research.google.com/drive/1a169o9gTP6GuaWp90R8QTrTZjtXM4eZb?usp=sharing>
4. Создание стека моделей URL:
<https://colab.research.google.com/drive/1dSnTC1yZNdorAJrjY9EJ7VjciXltWEQK?usp=sharing>
5. Сетевой анализ пользователей по специальности теплоэнергетика URL:
<https://colab.research.google.com/drive/1dSnTC1yZNdorAJrjY9EJ7VjciXltWEQK?usp=sharing>

ПРИЛОЖЕНИЕ В ОСНОВНЫЕ ФУНКЦИИ

Функция сбора групп

```
def groups_id(access_token, series):
    all_groups = []
    for user_id in series:
        time.sleep(1)
        groups = vk_api.groups.get(access_token=access_token,
                                    user_id= user_id,
                                    extended =1,
                                    count=10,
                                    fields = ['activity', 'description'],
                                    v=5.131)
        all_groups.extend(groups['items'])
    alist= []
    for i in range(len(series)): # Пробегаем циклом собираем группы по
10 штук
        for j in range(10):
            alist.append(i)
    try:
        for z,k in enumerate(alist):
            all_groups[z]['id'] = series[k] # Каждой группе
присваиваем id участника и его activity
    except:
        pass
    return all_groups [['id', 'description', 'activity', 'name',
'screen_name']]
```

Функция препроцессинга данных

```
def text_preprocess1(input_text):
    # Удалим все гиперссылки
    text = re.sub(r"https?:\/\/[^\s]+,?", "", input_text)
    # Удалим все цифры
    text = re.sub(r'\d', '', text)
    # Приведем к нижнему регистру
    text = text.lower()
    # Удалим все переносы
    text = text.replace('\n', ' ')
    # Удалили все смайлики
    text = re.sub('[^\x00-\x7Fa-яА-Я]', '', text)
    # Удалили все спецсимволы
    text = re.sub(r"[^a-zA-Z0-9a-яА-Я]", " ", text)

    # удалим id
    text = text.replace('id', '')
    # Удалили более одного пробела
    text = re.sub(r'\s+ ', ' ', text)
    tokenized_text = word_tokenize(text)
    clean_text = [word for word in tokenized_text if word not in filter]
    lemmatized_text = [parser.parse(word)[0].normal_form for word in
                        clean_text]
    lemmatized_text = ' '.join(lemmatized_text)

    return lemmatized_text
```

Алгоритм создания стека моделей

```
# Создаем пустой список
l=[]
# Делаем комбинацию данных
for df in combinations([algebra, lech_fiz, lech_delo, teploenrgetica], 2):
    l.append(shuffle(pd.concat(list(df),
axis=0)[features]).reset_index(drop=True))
# У нас получился список с нашими перемешанными датафреймами
# Создаем пустой список, чтобы хранить модели
l_models = []
for df_x in l:
    X_train, X_test, y_train, y_test = train_test_split(df_x.drop('target',
axis=1), df_x['target'], test_size=.3, random_state=7575)
    # Делаем пулы для подачи в модель
    train_pool = Pool(
        X_train, y_train,
        text_features=text_features
    )
    validation_pool = Pool(
        X_test, y_test,
        text_features=text_features
    )
    # Обучаем модель
    model_fin2 = fit_model_2(train_pool, validation_pool,
        dictionaries = [{
            'dictionary_id': 'Word',
            'max_dictionary_size': '50000'
        }], feature_calcers = ['BoW:top_tokens_count = 10000'])
    # Добавляем обученную модель в список моделей
    l_models.append(model_fin2)
```