Федеральное государственное образовательное бюджетное
учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»
(Финансовый университет)

Факультет информационных технологий и анализа больших данных
Кафедра искусственного интеллекта

Выпускная квалификационная работа

на тему: «Применение методов машинного обучения для выявления англицизмов в русскоязычных новостных статьях»

Направление подготовки: 01.04.02 Прикладная математика и информатика

Направленность программы: «Машинное обучение на текстах и графах»

Выполнил студент учебной группы
МО23-1м
Гордеева Наталия Георгиевна

Руководитель к.т.н., доцент
К.А. Маковейчук

ВКР соответствует предъявляемым требованиям
Заведующий кафедрой искусственного интеллекта, к.э.н.

_____М.В. Коротеев

«___» _____2025 г.

Москва 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В современном мире процессы глобализации и стремительного развития информационных технологий оказывают значительное влияние на все аспекты жизни общества, включая язык как основное средство коммуникации. Особенно заметным становится влияние английского языка на другие языки мира, в том числе на русский язык. Это влияние проявляется прежде всего через заимствования английских слов и выражений – англицизмов, которые активно проникают в русскоязычное информационное пространство.

Средства массовой информации, и, в частности, новостные издания, являются одним из основных каналов распространения англицизмов в русском языке. Журналисты и редакторы часто используют английские заимствования для описания новых явлений в технологической, экономической и социальной сферах, где русскоязычные эквиваленты либо отсутствуют, либо являются менее точными или удобными в использовании. Такие слова как "стартап", "дедлайн", "тренд", "контент" прочно вошли в лексикон современных СМИ и активно используются в новостных статьях.

Актуальность исследования обусловлена несколькими факторами. Во-первых, растущим количеством англицизмов в русскоязычных текстах, что создает необходимость их систематического изучения и анализа. Во-вторых, потребностью в автоматизации процесса выявления и анализа заимствований для более эффективного мониторинга языковых изменений. В-третьих, развитием методов машинного обучения и их успешным применением в задачах обработки естественного языка открывает новые возможности для автоматического анализа лингвистических явлений, включая заимствования.

Современные методы машинного обучения и, в частности, глубокого обучения, демонстрируют впечатляющие результаты в различных задачах обработки естественного языка. Однако их применение для выявления англицизмов в русскоязычных текстах остается малоизученной областью, что создает простор для исследований и разработки новых подходов.

Целью данной работы является разработка и исследование методов машинного обучения для автоматического выявления англицизмов в русскоязычных новостных статьях.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ существующих подходов к выявлению заимствований в текстах и определить их преимущества и недостатки.
2. Исследовать лингвистические характеристики англицизмов в русскоязычных новостных текстах.
3. Разработать методику применения алгоритмов машинного обучения для автоматического выявления англицизмов.
4. Создать и разметить корпус текстов для обучения и тестирования моделей.
5. Реализовать и сравнить различные методы машинного обучения для решения поставленной задачи.
6. Провести экспериментальную оценку разработанных методов и проанализировать полученные результаты.

Объектом исследования являются англицизмы в современных русскоязычных новостных текстах.

Предметом исследования выступают методы машинного обучения для автоматического выявления англицизмов в текстах.

В работе используются следующие методы исследования:

- Методы корпусной лингвистики для сбора и анализа текстовых данных;
- -Методы компьютерной лингвистики для предварительной обработки текстов;
- Методы машинного обучения, включая алгоритмы классификации и глубокие нейронные сети;
- Статистические методы для оценки результатов экспериментов;

- Сравнительный анализ для оценки эффективности различных подходов.

Теоретическая значимость работы заключается в систематизации знаний о применении методов машинного обучения для анализа лексических заимствований и разработке новых подходов к автоматическому выявлению англицизмов в текстах. Практическая значимость состоит в создании инструментария для автоматического анализа заимствований в новостных текстах, который может быть использован лингвистами, редакторами и исследователями языка для мониторинга языковых изменений и анализа тенденций заимствования.

Научная новизна исследования определяется следующими факторами:

1. Разработкой гибридного метода автоматического выявления англицизмов на русском языке;

2. Созданием размеченного корпуса текстов для обучения моделей выявления англицизмов;

3. Использование семейства моделей Qwen для решения данной задачи.

Структура работы включает введение, три основных раздела, заключение, список использованных источников и приложения. В первом разделе проводится анализ методов машинного и глубокого обучения для выявления англицизмов в текстах, рассматриваются лингвистические характеристики англицизмов и существующие инструменты для обработки русскоязычных текстов. Второй раздел посвящен разработке методики применения машинного обучения для выявления англицизмов и описанию процесса создания обучающего корпуса. В третьем разделе представлены результаты экспериментальной оценки разработанных методов и их анализ.

Результаты исследования могут найти применение в различных областях, связанных с анализом текстов и языковых изменений, включая: лингвистические исследования динамики заимствований, разработку систем автоматической обработки текстов, создание инструментов для редакторской работы, мониторинг языковых изменений в СМИ.

Следовательно, данное исследование направлено на решение актуальной задачи автоматического выявления англицизмов в русскоязычных текстах с использованием современных методов машинного обучения и имеет как теоретическую, так и практическую значимость для развития компьютерной лингвистики и обработки естественного языка.

1. Анализ методов машинного и глубокого обучения для выявления англицизмов в текстах

1.1. Лингвистический анализ англицизмов и их характеристика

Англицизмы, как лексический феномен, представляют собой многогранный лингвистический феномен, характеризующийся интеграцией слов и выражений англоязычного происхождения в систему русского языка с последующей адаптацией к его фонетическим, морфологическим и грамматическим нормам. Современная тенденция глобализационных процессов и возрастающее влияние англоязычной культуры во всех сферах жизнедеятельности привели к формированию значительного пласта заимствованной лексики, функционирующей в различных коммуникативных сферах, включая массмедийное пространство, деловую коммуникацию, информационные технологии, научно-образовательную среду. Новостной дискурс, выступающий одним из основных каналов массовой информации, активно использует англицизмы для номинации инновационных концептов, технологических разработок и социокультурных явлений, существенно влияя на формирование языковой картины мира русскоговорящего населения.

Исследовательские разработки Пульчини и других авторитетных лингвистов позволяют дифференцировать несколько категорий лексических заимствований, характерных для новостного дискурса. Фразовые заимствования, представляющие собой устойчивые словесные комплексы, наиболее часто встречаются в материалах делового и технологического характера, обеспечивая терминологическую точность при описании специализированных областей. Прямые лексические заимствования сохраняют структурное и семантическое сходство с англоязычным прототипом, функционируя как на уровне отдельных лексем, так и в составе многокомпонентных единиц, что создает определенные закономерности в процессе их адаптации к системе русского языка.

Адаптационные механизмы англицизмов в русскоязычном пространстве характеризуются процессом лингвистической доместикации – комплексом фонетических, грамматических и семантических трансформаций, направленных на интеграцию иноязычного материала в принимающую языковую систему. Специфика новостных текстов проявляется в сохранении более тесной связи заимствований с исходными формами английского языка, что обусловлено стремлением к терминологической точности и созданию эффекта международной компетентности, однако подобная практика существенно усложняет процессы автоматической обработки текстовых массивов и требует разработки специализированных алгоритмических подходов.

Морфологическая адаптация заимствованной лексики реализуется преимущественно через механизмы суффиксации, среди которых наибольшую продуктивность демонстрируют формы с элементами "-ер" (спикер, блогер), "-инг" (консалтинг, маркетинг), "-мент" (менеджмент, истеблишмент) [25]. Системность в образовании производных форм представляет значительный интерес для разработки компьютерных алгоритмов, ориентированных на автоматическое распознавание и классификацию англицизмов в корпусах русскоязычных текстов различной жанровой принадлежности.

Функциональная специфика англицизмов в новостном дискурсе проявляется в нескольких аспектах. Заполнение лексических лакун становится первостепенной задачей при обозначении инновационных концептов и феноменов, не имеющих устоявшихся русскоязычных эквивалентов. Стилистическая функция реализуется через создание эффекта современности и международного контекста, что повышает степень доверия аудитории к представляемой информации и формирует имидж издания как авторитетного источника актуальных новостей мирового масштаба.

Аналитический подход к исследованию новостных текстов требует учета феномена псевдоанглицизмов – лексических единиц, формально

напоминающих английские заимствования, но обладающих семантическими характеристиками, отличными от оригинального значения или вовсе отсутствующих в англоязычном лексиконе. Рассматриваемое явление создает дополнительные трудности при разработке и применении автоматизированных систем анализа текстового материала, требуя имплементации более сложных алгоритмов семантической дифференциации.

Классификационные модели англицизмов в русском языке, предложенные Г. Андерсеном, включают несколько основных категорий: прямые заимствования с сохранением исходных форм и значений (блогер, стартап), гибридные конструкции с комбинацией английских и русских морфем (овербукинг, фейсконтроль), кальки с дословным переводом англоязычных выражений (утечка мозгов, мыльный пузырь) и псевдоанглицизмы со специфическим семантическим наполнением (смокинг в значении "вечерний костюм") [9]. Подобное разнообразие адаптационных форм свидетельствует о сложности процессов языкового взаимодействия и требует многоаспектного подхода при разработке методик автоматического распознавания заимствований.

Лингвистическая характеристика англицизмов в русском языке определяется их адаптацией к фонетической, морфологической и синтаксической системе языка-реципиента. Исследования М. Горлаха хбъсвидетельствуют о комплексном характере адаптационных процессов, включающих транслитерацию, транскрипцию и морфологическую трансформацию. Суффиксальные модели "-инг" (консалтинг), "-ер" (блогер), "-мент" (эндорсмент) способствуют натурализации иноязычного материала в русскоязычной среде, однако сохранение элементов иностранного происхождения может создавать коммуникативные барьеры для отдельных категорий носителей языка, особенно представителей старшего поколения или лиц, не связанных с соответствующими профессиональными областями.

Технологическая, финансовая, спортивная и культурная сферы демонстрируют наиболее высокую концентрацию англицизмов в новостном

дискурсе. Информационно-технологическая терминология включает такие единицы, как блокчейн, криптовалюта, искусственный интеллект, финансовая сфера характеризуется использованием терминов дефолт, хеджирование, инвестиция, спортивная журналистика оперирует понятиями трансфер, драфт, чемпионшип. Применение англицизмов в новостных публикациях позволяет авторам с максимальной точностью отражать современные реалии и тенденции, хотя избыточное использование заимствованной лексики может негативно влиять на восприятие информации широкой аудиторией.

Исследовательские работы А. Феногеновой и соавторов [5] акцентируют внимание на проблематике автоматического анализа англицизмов в русскоязычных текстах, рассматривая возможности идентификации и замены заимствований на исконно русские эквиваленты для повышения читабельности материала. Методы машинного обучения, основанные на анализе фонетических, морфологических и синтаксических параметров заимствованных слов, представляют значительный потенциал для разработки эффективных алгоритмов распознавания англицизмов, учитывая тенденцию к сохранению оригинальных корневых морфем при заимствовании.

Научные разработки Е. Мелладо и К. Лигноса [12] в области автоматического обнаружения неассимилированных заимствований в испанском языке предлагают методологические подходы, применимые к анализу англицизмов в русском языке. Использование аннотированных текстовых корпусов для обучения моделей машинного анализа позволяет идентифицировать заимствования на основе их лингвистических характеристик с учетом контекстуальных особенностей употребления, влияющих на восприятие и интерпретацию языкового материала.

Исследования Д. Граддола [2] относительно влияния английского языка на другие языковые системы в условиях глобализационных процессов характеризуют английский язык как глобального экспортера лексического материала, что объясняет масштабное распространение англицизмов в различных языках мира, включая русский. Особая роль заимствований в

обозначении инновационных технологий и концепций подтверждает их значимость в формировании современного языкового ландшафта и определяет необходимость всестороннего изучения данного лингвистического феномена.

Англицизмы занимают особое положение в языке новостной журналистики, выполняя ряд специфических функций в контексте современного медийного пространства, где оперативность информационного обмена и терминологическая точность приобретают первостепенное значение. Новостные публикации, выступающие ключевым источником информации для широких слоев населения, активно используют англоязычные заимствования при описании технологических инноваций, экономических процессов, культурных феноменов и социальных тенденций, что связано с первичным возникновением соответствующих концептов и терминов в англоязычной среде с последующим заимствованием другими языковыми системами.

Терминологическая точность и актуальность представляют собой первостепенные функциональные характеристики англицизмов, позволяющие журналистам и авторам новостных материалов максимально адекватно отражать современную действительность и инновационные тенденции в различных сферах общественной жизни. Технологические термины блокчейн, криптовалюта, искусственный интеллект стали неотъемлемыми элементами терминологического аппарата публикаций соответствующей тематики, компенсируя отсутствие точных русскоязычных эквивалентов для обозначения новых феноменов цифровой эпохи.

Стилистический потенциал англицизмов активно реализуется в публикациях, посвященных молодежной культуре, модной индустрии и современным трендам социального развития. Лексические единицы тренд, хайп, лайфстайл способствуют формированию динамичного и современного стилистического облика текста, соответствующего ожиданиям целевой аудитории и подчеркивающего актуальность представляемой информации в контексте глобальных культурных процессов (таблица 1.1.1).

Таблица 1.1.1

Классификация англицизмов по лингвистическим характеристикам

Группа	Примеры англицизмов	Описание
Прямые заимствования	блогер (blogger), стартап (startup), дефолт (default), хеджирование (hedging)	Слова, которые сохраняют свою оригинальную форму и значение.
Гибридные формы	овербукинг (overbooking), фейсконтроль (face control), онлайн-магазин (online store)	Сочетание английских и русских морфем.
Кальки	утечка мозгов (brain drain), мыльный пузырь (soap bubble), голубой экран (blue screen)	Слова или выражения, которые переводятся дословно с английского языка.
Псевдоанглицизмы	смокинг (smoking), автостоп (autostop), паркинг (parking)	Слова, которые внешне напоминают английские, но имеют иное значение.
Технические термины	блокчейн (blockchain), криптовалюта (cryptocurrency), искусственный интеллект (AI)	Термины, связанные с технологиями и наукой.
Экономические термины	инвестиция (investment), дефолт (default), хеджирование (hedging)	Термины, используемые в финансовой и экономической сферах.
Спортивные термины	трансфер (transfer), драфт (draft), чемпионшип (championship)	Термины, связанные со спортом.
Культурные термины	тренд (trend), хайп (hype), лайфстайл (lifestyle)	Термины, связанные с культурой, модой и общественными трендами.

Источник: составлено автором

В контексте лингвистического исследования англоязычных заимствований в русскоязычных текстовых массивах, в частности в новостном дискурсе, целесообразно применить многоаспектную классификацию лексических единиц на основании их структурно-семантических характеристик и функциональной специфики. Методологический фундамент данной типологии составляют научные разработки М. Горлаха [6], А. Дьякова [24], а также исследовательские проекты А. Феногеновой [5] и группы соавторов, сформировавшие системный подход к анализу иноязычных элементов в русскоязычном коммуникативном пространстве.

Категория прямых заимствований охватывает лексические единицы, сохраняющие структурно-семантическое сходство с исходными

англоязычными прототипами при минимальной адаптации к принимающей языковой системе. Термины блогер, стартап, дефолт, хеджирование регулярно фиксируются в новостных материалах при описании инновационных технологических разработок, современных экономических механизмов и социальных феноменов глобального характера. Отсутствие точных русскоязычных эквивалентов для номинации соответствующих концептов определяет необходимость обращения к англоязычным заимствованиям, обеспечивающим максимальную информационную точность при минимальных затратах языковых ресурсов.

Гибридные образования характеризуются комбинаторикой английских и русских морфологических элементов, демонстрируя высокую степень лингвистической креативности при интеграции иноязычного материала в систему русского языка. Лексемы овербукинг, фейсконтроль, онлайн-магазин широко представлены в новостных публикациях технологической, бизнес-ориентированной и потребительской тематики, обеспечивая оптимальный баланс между терминологической точностью и лингвистической натурализацией при сохранении семантической связи с оригинальными англоязычными терминами.

Кальки представляют собой словесные комплексы, созданные путем дословного перевода англоязычных выражений с сохранением структурно-семантических характеристик исходных единиц. Выражения утечка мозгов, мыльный пузырь, голубой экран регулярно используются в новостных материалах для описания комплексных явлений и процессов, не имеющих прямых соответствий в русском языке. Калькирование позволяет сохранить семантическую целостность оригинального концепта при полной адаптации его вербального выражения к нормам русского языка, что существенно облегчает восприятие информации широкой аудиторией при сохранении терминологической точности.

Псевдоанглицизмы формируют особый класс лексических единиц, внешне напоминающих англоязычные заимствования, но обладающих

семантическими характеристиками, отличными от оригинальных значений или функционирующих в контекстах, нетипичных для английского языка. Термины смокинг в значении "вечерний костюм", автостоп как "путешествие на попутных машинах", паркинг в смысле "стоянка для автомобилей" встречаются преимущественно в публикациях модной, культурной и общественной тематики, создавая эффект современности и интернациональности при сохранении доступности материала для русскоязычной аудитории.

Лингвистический анализ англицизмов в русскоязычных текстовых корпусах позволяет выявить системные закономерности адаптации иноязычного материала к нормам принимающего языка при сохранении определенных оригинальных характеристик заимствованных единиц. Функционирование англицизмов в новостном дискурсе отражает глобальные тенденции информационного обмена и современные реалии социокультурного развития, требуя при этом комплексного подхода к идентификации заимствований и оценке возможностей их замены русскоязычными эквивалентами для обеспечения максимальной доступности текстового материала различным категориям читательской аудитории. Последующие разделы исследования будут посвящены методологическим аспектам автоматического анализа текстовых массивов с целью выявления англицизмов, а также обзору существующих технологических решений и инструментальных средств для обработки русскоязычных текстов различной жанровой принадлежности.

1.2. Методы автоматического анализа текста и выявления заимствований

Автоматизированный анализ текстовых массивов с целью идентификации лексических заимствований, в частности англицизмов, представляет собой актуальное направление исследований в области

компьютерной лингвистики и обработки естественного языка. Современные глобализационные процессы и возрастающее влияние англоязычной культуры в различных сферах социальной жизни обуславливают необходимость разработки эффективных алгоритмических подходов к выявлению и обработке заимствованной лексики для оптимизации процессов автоматического анализа текстовой информации. Методологический арсенал данной области включает как традиционные подходы, основанные на лингвистических правилах и лексикографических ресурсах, так и инновационные технологии машинного обучения и глубинных нейронных сетей, обеспечивающие комплексный анализ языкового материала с учетом контекстуальных особенностей его функционирования.

Классические методы автоматической идентификации заимствований базируются на применении лингвистических правил и специализированных словарей, содержащих репрезентативные наборы англицизмов различных категорий. Подобные подходы предполагают систематизацию фонетических, морфологических и синтаксических характеристик заимствованной лексики для формирования формализованных критериев ее распознавания в текстовых корпусах различной жанровой принадлежности.

Правило-ориентированный анализ представляет собой методологический подход, основанный на формализации лингвистических закономерностей адаптации иноязычного материала к системе русского языка. Систематизация структурных маркеров заимствований, таких как характерные суффиксальные комплексы "-инг" (консалтинг), "-ер" (блогер), "-мент" (эндорсмент), позволяет разработать алгоритмические правила идентификации соответствующих единиц в текстовых массивах. Исследования М. Горлаха [6] свидетельствуют об определенных ограничениях данного подхода, связанных с наличием нестандартных форм адаптации заимствований и значительной вариативностью их контекстуального употребления, что существенно затрудняет создание универсальных правил

распознавания англицизмов в текстах различной стилистической направленности.

Лексикографический подход основывается на использовании специализированных словарей, содержащих комплексные перечни англоязычных заимствований с указанием их фонетических, морфологических и семантических характеристик, а также русскоязычных эквивалентов соответствующих единиц. Подобный метод обеспечивает высокую скорость идентификации известных англицизмов в текстовых массивах, однако характеризуется существенными методологическими ограничениями. Динамический характер процессов лексического заимствования приводит к неизбежной неполноте словарных ресурсов, не отражающих в полной мере состав новейших англицизмов, активно функционирующих в современной коммуникативной среде. Контекстуальная специфика употребления заимствованной лексики также не получает адекватного отражения в лексикографических источниках, что существенно ограничивает эффективность данного подхода при анализе текстов с высокой степенью семантической вариативности.

Современный этап развития компьютерной лингвистики характеризуется значительным прогрессом в области машинного обучения и нейросетевых технологий, обеспечивающих качественно новый уровень автоматического анализа текстовой информации. Инновационные методологические подходы позволяют учитывать не только формальные характеристики лексических единиц, но и широкий спектр контекстуальных параметров их функционирования, что существенно повышает точность идентификации англицизмов различных категорий в текстовых корпусах разнообразной жанровой принадлежности.

Классификационные алгоритмы машинного обучения представляют собой эффективный инструментарий для идентификации англицизмов на основе анализа аннотированных текстовых корпусов. Исследовательские проекты А. Феногеновой и группы соавторов [5] демонстрируют

продуктивность применения методов машинного обучения для выявления англоязычных заимствований в русскоязычных текстах на основе анализа структурных особенностей соответствующих лексических единиц. Алгоритмические подходы, основанные на сравнении с англоязычными прототипами, демонстрируют высокую эффективность при использовании таких классификационных моделей, как Support Vector Machines и Random Forest, обучаемых на специализированных корпусах, содержащих репрезентативные выборки англицизмов в различных контекстуальных окружениях.

Нейросетевые технологии обеспечивают возможность моделирования комплексных зависимостей между лексическими единицами и их контекстуальным окружением на основе многоуровневого анализа текстовой информации. Научные разработки Е. Мелладо и К. Лигноса [12] в области автоматической идентификации неассимилированных заимствований в испанском языке демонстрируют эффективность применения глубинных нейронных сетей для решения аналогичных задач в русскоязычной среде. Рекуррентные нейронные сети и трансформерные архитектуры, обучаемые на объемных корпусах аннотированных текстов, демонстрируют высокие показатели точности при идентификации заимствований различных категорий в разнообразных контекстуальных условиях.

Трансформерные модели, такие как BERT и GPT, представляют собой передовые архитектурные решения в области обработки естественного языка, обеспечивающие комплексный анализ контекстуальных особенностей функционирования лексических единиц в текстовых массивах различной стилистической направленности. Исследовательские проекты Д. Лукичева и группы соавторов [11] демонстрируют высокую эффективность применения трансформерных моделей ruRoberta-large и XLM-RoBERTa для выявления и замены англицизмов в русскоязычных текстах, особенно в контексте новостных публикаций, характеризующихся высокой степенью семантической вариативности и терминологической насыщенности.

Методологические инновации в области настройки предобученных языковых моделей, такие как prompt-tuning и Low-Rank Adaptation, обеспечивают существенное повышение эффективности автоматического анализа текстовой информации при минимизации вычислительных затрат. Адаптация крупных языковых моделей для решения специализированных задач по идентификации англицизмов в текстовых корпусах приобретает особую значимость в контексте обработки масштабных массивов новостной информации, требующей оперативного анализа и классификации.

Метод Low-Rank Adaptation демонстрирует значительный потенциал в области оптимизации процессов настройки языковых моделей для специализированных задач компьютерной лингвистики. Сокращение количества обучаемых параметров при сохранении высоких показателей производительности достигается за счет имплементации малоранговых матричных структур в архитектуру предобученных моделей, что обеспечивает эффективную адаптацию к конкретным задачам анализа текстовой информации при существенном снижении вычислительных затрат на обучение и применение соответствующих алгоритмических решений.

Детальный анализ методологических подходов к автоматической идентификации англицизмов в русскоязычных текстах позволяет провести сравнительную оценку эффективности традиционных и современных алгоритмических решений на основе систематизации их ключевых характеристик, применяемых корпусов, исследовательских проектов и достигнутых результатов, представленных в специализированных публикациях (Приложение А).

Правило-ориентированный анализ демонстрирует высокую эффективность при идентификации англицизмов с четко выраженными формальными признаками, такими как характерные суффиксальные комплексы и корневые морфемы. Данный подход, однако, характеризуется существенным снижением результативности при работе с нестандартными формами и сложными контекстуальными окружениями, требующими более

гибкой методологической организации процессов анализа. Исследовательские проекты А. Феногеновой и соавторов [5] демонстрируют возможность применения правило-ориентированного подхода для предварительной фильтрации англицизмов, при этом точностные показатели данного метода составляют 60-70% для стандартных случаев с последующим существенным снижением при усложнении анализируемого материала.

Лексикографический подход обеспечивает оперативную идентификацию зафиксированных англицизмов, получивших отражение в специализированных словарных ресурсах. Основные ограничения данного метода связаны с неполнотой существующих лексикографических источников и их принципиальной неспособностью охватить новейшие заимствования, активно функционирующие в современной коммуникативной среде. Исследования А. Дьякова [24] подтверждают продуктивность применения словарного подхода для анализа устоявшихся англицизмов при его ограниченной эффективности в контексте работы с новейшими и редкими заимствованиями. Точностные показатели данного метода составляют 70-80% для известных англицизмов с существенным снижением при анализе новейших заимствований, не получивших отражения в лексикографических источниках.

Алгоритмы машинного обучения, включая SVM и Random Forest, демонстрируют высокую эффективность при учете контекстуальных параметров функционирования англицизмов и обработке новых форм заимствованной лексики. Данные методы обеспечивают высокие точностные показатели при наличии репрезентативных корпусов аннотированных текстов, однако характеризуются снижением эффективности в условиях ограниченности данных или сложных контекстуальных окружений. Исследовательские проекты А. Феногеновой и соавторов [5] свидетельствуют о возможности применения SVM для выявления англицизмов в текстах социальных медиа при условии тщательной настройки параметров

соответствующих моделей. Точностные показатели данного метода составляют 75-85% при наличии достаточного объема обучающих данных.

Рекуррентные нейронные сети, включая архитектуры LSTM и GRU, обеспечивают эффективную обработку протяженных последовательностей и комплексный учет контекстуальных параметров функционирования лексических единиц. Данные модели демонстрируют высокие результаты в задачах классификации и генерации текстовой информации, однако характеризуются значительными вычислительными затратами и сниженной интерпретируемостью механизмов функционирования. Исследовательские проекты Мелладо и Лигноса [12] подтверждают эффективность применения рекуррентных нейронных сетей для идентификации заимствований в испанском языке при условии формирования объемных корпусов обучающих данных. Точностные показатели данного метода составляют 80-90% при анализе протяженных текстовых массивов, однако требуют значительных вычислительных ресурсов для практической реализации.

Трансформерные модели, включая архитектуры BERT и GPT, представляют собой наиболее эффективные инструменты автоматического анализа текстовой информации и идентификации лексических заимствований в современной компьютерной лингвистике. Данные архитектурные решения обеспечивают моделирование комплексных зависимостей между лексическими единицами и их контекстуальным окружением, что делает их особенно продуктивными при анализе новостных публикаций с высокой степенью семантической вариативности. Исследовательские проекты Д. Лукичева и соавторов [11] свидетельствуют о высокой эффективности моделей ruRoberta-large и XLM-RoBERTa при идентификации англицизмов в контексте новостных публикаций, несмотря на значительные требования к объему обучающих данных и вычислительным ресурсам. Точностные показатели данного метода составляют 90-95% при анализе сложных контекстуальных окружений, что существенно превосходит результативность других методологических подходов в аналогичных условиях.

В контексте задачи идентификации англицизмов в русскоязычных новостных публикациях наиболее эффективными инструментами представляются трансформерные модели ruRoberta-large и XLM-RoBERTa, обеспечивающие комплексный учет контекстуальных параметров функционирования лексических единиц и высокие точностные показатели при решении задач классификации и генерации текстовой информации. Данные архитектурные решения демонстрируют максимальную эффективность при анализе новостных публикаций, характеризующихся значительной ролью контекстуальных параметров в определении семантического наполнения лексических единиц.

Альтернативным методологическим подходом может выступать применение алгоритмов машинного обучения, включая SVM и Random Forest, для предварительной фильтрации англицизмов в текстовых массивах, особенно в условиях ограниченности доступных данных для обучения моделей. Максимальные показатели точности и гибкости при идентификации англицизмов различных категорий в разнообразных контекстуальных окружениях достигаются при использовании трансформерных архитектур, демонстрирующих наиболее высокую эффективность в современной компьютерной лингвистике.

1.3. Анализ существующих решений и инструментов для обработки русскоязычных текстов

Программно-инструментальное обеспечение обработки русскоязычных текстовых массивов характеризуется значительным разнообразием специализированных библиотек и моделей, ориентированных на учет лингвоспецифических особенностей русского языка, таких как морфологическая вариативность, синтаксическая гибкость и развитая словообразовательная система. Современный арсенал технологических решений охватывает как универсальные библиотеки обработки естественного

языка, так и узкоспециализированные инструменты для решения конкретных лингвистических задач, включая идентификацию лексических заимствований, в частности англицизмов, в текстах различной жанровой принадлежности.

Библиотека SpaCy представляет собой высокопроизводительный инструментарий обработки естественного языка, характеризующийся оптимизированной архитектурой и эффективными алгоритмическими решениями для анализа масштабных текстовых корпусов. Ключевым преимуществом данного инструмента выступает его способность к оперативной обработке значительных объемов текстовой информации, что подтверждается экспериментальными данными о производительности: средние временные затраты на анализ 10000 лексических единиц составляют всего 3.3473E-03 секунды, обеспечивая существенное преимущество в скорости по сравнению с альтернативными решениями. Интеграционные возможности SpaCy с архитектурами глубокого обучения расширяют функциональный потенциал библиотеки, позволяя применять ее для решения комплексных задач классификации текстов и извлечения информации [51]. Ограниченная поддержка русского языка по сравнению с англоязычным сегментом требует дополнительной настройки и дообучения моделей для оптимизации результатов при работе с русскоязычными текстовыми массивами. Результаты показали, что spaCy демонстрирует высокую скорость обработки текстов, что подтверждается данными из сравнительной таблицы 1.3.1:

Таблица 1.3.1

Объем обрабатываемых данных (лексем) за раличное среднее время выполнения (с) для spaCy

Наименование функции	Объем обрабатываемых данных (лексем)	Среднее время выполнения (с)
spaCy	100	4.6847E-05
	1000	4.2162E-04
	2500	1.0031E-03
	5000	1.9003E-03
	10000	3.3473E-03

Источник: данные статьи К.С. Макарова, А.А. Халина «Сравнительный анализ библиотек для обработки естественного языка» [31]

Библиотека Natural Language Toolkit представляет собой один из старейших и наиболее комплексных инструментариев лингвистического анализа, предоставляющий разработчикам широкий спектр функциональных возможностей для решения разнообразных задач компьютерной лингвистики. Относительно низкие показатели производительности (обработка 10000 лексических единиц занимает 2.3383 секунды) компенсируются богатством аналитического инструментария, включающего модули токенизации, лемматизации, морфологического и синтаксического анализа текстовой информации. Расширенная поддержка морфологического анализа русскоязычных текстов делает NLTK особенно ценным инструментом при решении задач идентификации заимствованной лексики, требующих комплексного учета морфологических характеристик анализируемых единиц [43].

Natasha выступает специализированной библиотекой для обработки русскоязычных текстов, разработанной с учетом лингвоспецифических особенностей русского языка и оптимизированной для решения широкого спектра задач компьютерной лингвистики в русскоязычном сегменте [37]. Функциональный арсенал библиотеки включает модули морфологического анализа, синтаксического разбора и распознавания именованных сущностей, учитывающие структурно-семантические характеристики русского языка. В научной статье Г.О. Сидорова [33] автор описывает применение Natasha для извлечения именованных сущностей из русскоязычных текстов. Результаты показывают, что Natasha демонстрирует высокую точность при работе с русскоязычными данными, что делает её особенно полезной для задач, связанных с анализом новостных статей, где важно идентифицировать ключевые элементы текста, такие как имена людей, организаций и географических объектов.

Векторные представления лексических единиц русского языка реализуются в нескольких высокоэффективных моделях, демонстрирующих различные соотношения качественных характеристик и ресурсоемкости. Navec представляет собой компактную и производительную модель, обеспечивающую высокие качественные показатели (0.719) при решении задач определения семантической близости лексических единиц при относительно невысоких требованиях к объему оперативной памяти (50.6 МБ). RusVectores предлагает линейку моделей с вариативными характеристиками, демонстрирующих качественные показатели в диапазоне 0.638–0.726 при более высоких требованиях к ресурсному обеспечению (220.6–290.7 МБ), что позволяет выбирать оптимальные параметры модели в зависимости от специфики решаемых задач и доступных вычислительных ресурсов.

Современный этап развития компьютерной лингвистики характеризуется доминированием предобученных трансформерных моделей, таких как BERT, GPT и T5, демонстрирующих высокую эффективность при решении широкого спектра задач обработки естественного языка. Архитектурные особенности данных моделей обеспечивают комплексный учет контекстуальных параметров и семантических связей между лексическими единицами, что существенно повышает эффективность процессов анализа текстовой информации. Русскоязычный сегмент представлен такими моделями, как ruBERT, ruGPT и ruT5, адаптированными к специфике русского языка и применимыми для идентификации англицизмов в текстовых массивах различной жанровой принадлежности [36].

FRED-T5 представляет собой инновационное семейство моделей с архитектурой энкодер-декодер, демонстрирующее высокую эффективность при решении задач преобразования текстовой информации. Вариативность параметрического объема модели (1.74B и 820M параметров) позволяет выбирать оптимальный баланс между производительностью и ресурсоемкостью в зависимости от конкретных условий применения и

доступных вычислительных ресурсов, что делает данное семейство моделей гибким инструментом для решения разнообразных задач компьютерной лингвистики.

RuBERT от компании AI-Forever демонстрирует впечатляющие результаты при решении задач классификации и анализа текстовых массивов на русском языке. Обучение модели на масштабном корпусе русскоязычных текстов объемом 30 гигабайт обеспечивает высокую точность анализа в различных контекстуальных условиях. Вариативность параметрического объема модели (base с 178M параметров и large с 427M параметров) позволяет выбирать оптимальную конфигурацию для конкретных задач с учетом доступных вычислительных ресурсов. Высокая эффективность RuBERT при решении задач классификации делает данную модель особенно ценным инструментом для идентификации заимствованной лексики в текстовых массивах различной жанровой принадлежности [48].

RuRoBERTa-large представляет собой усовершенствованную версию BERT-подобной архитектуры, обученную на значительно большем объеме данных (250 гигабайт русскоязычных текстов), что обеспечивает существенное повышение эффективности при решении различных задач компьютерной лингвистики. Параметрический объем модели (355M параметров) позволяет достигать превосходных результатов при решении задач классификации и сравнения текстовых массивов, что делает RuRoBERTa-large оптимальным инструментом для идентификации англицизмов в русскоязычных текстах различной стилистической направленности.

Практический опыт применения рассмотренных инструментов в исследовательских проектах по идентификации англицизмов демонстрирует их значительный потенциал для решения соответствующих задач компьютерной лингвистики. Исследования морфологической адаптации англицизмов в русском языке с использованием SpaCy показали высокую эффективность данной библиотеки при выявлении заимствованных

лексических единиц на основе их морфологических характеристик, что подтверждает продуктивность применения данного инструмента для решения задач идентификации заимствований в русскоязычных текстах. Проекты по автоматическому распознаванию англицизмов с использованием RuBERT продемонстрировали впечатляющие показатели точности (выше 90%) на тестовых наборах данных, что свидетельствует о высоком потенциале трансформерных моделей для решения задач идентификации заимствованной лексики в текстовых массивах различной жанровой принадлежности.

Сравнительный анализ различных моделей векторного представления лексических единиц позволяет определить оптимальные инструменты для решения конкретных задач компьютерной лингвистики с учетом специфики анализируемого материала и доступных вычислительных ресурсов, что имеет критическое значение при разработке эффективных алгоритмических решений для идентификации англицизмов в русскоязычных текстах (таблица 1.3.2):

Таблица 1.2.2

Сравнение качественных и количественных характеристик различных моделей векторного представления слов

Модель	Среднее качество на 6 датасетах	Время загрузки (с)	Размер модели (МБ)	Размер словаря ($\times 10^3$)
Navec hudlit_12B_500K_300d_100q	0.719	1.0	50.6	500
Navec news_1B_250K_300d_100q	0.653	0.5	25.4	250
RusVectores ruscorpora_upos_cbow_300_20_2019	0.692	3.3	220.6	189
RusVectores ruwikiruscorpora_upos_skipgram_300_2_2019	0.691	5.0	290.0	248
RusVectores tayga_upos_skipgram_300_2_2019	0.726	5.2	290.7	249

Источник: Habr.com

Для решения задачи идентификации англицизмов в русскоязычных новостных публикациях оптимальным представляется интегративный

методологический подход, базирующийся на синергии современных достижений компьютерной лингвистики и технологий глубокого обучения. Систематический анализ существующих исследовательских проектов и их практических результатов позволяет утверждать, что наивысшая эффективность достигается при комбинированном применении трансформерных архитектур, в частности RuBERT и RuRoBERTa-large, с алгоритмами морфологического анализа, специализированными для русскоязычных текстов.

Исследовательский проект А. Кутузова и коллектива соавторов [28], реализованный в 2021 году, предоставляет убедительные экспериментальные доказательства доминирующего положения архитектуры RuRoBERTa-large в задачах классификации заимствованной лексики с достижением беспрецедентных показателей точности (93.2%) при анализе новостных текстовых массивов. Исключительная эффективность данной модели обусловлена архитектурным совершенством механизма внимания, обеспечивающего глубинное понимание контекстуальных параметров функционирования лексических единиц, а также высокой результативностью алгоритмов подтокенизации, позволяющих идентифицировать новейшие англицизмы, еще не получившие широкого распространения в языковом узусе. Принципиально важным фактором выступает оптимизация модели для работы с русскоязычным материалом, что существенно повышает ее эффективность при решении задач идентификации англицизмов в текстах соответствующей языковой принадлежности.

Морфологический анализ с применением специализированной библиотеки Natasha представляет собой значимый компонент комплексного подхода к идентификации англицизмов в русскоязычных текстах. Научные разработки Ю.М. Соловьевой и группы соавторов [34], опубликованные в 2023 году, демонстрируют существенное повышение точности выявления заимствований (на 7.5%) при интеграции алгоритмов морфологического анализа в общую архитектуру аналитической системы. Значительный прирост

эффективности обусловлен специфической оптимизацией библиотеки для работы с морфологическими особенностями русского языка и высокой результативностью обработки различных аспектов словоизменения и словообразования, критически важных при идентификации адаптированных заимствований в русскоязычных текстах.

Масштабное исследование "Автоматическое выявление англицизмов в русскоязычных медиа" под руководством А.С. Петрова, реализованное в 2022 году, базировалось на применении трансформерной модели RuBERT для анализа корпуса новостных публикаций объемом 50000 текстовых единиц. Интеграция алгоритмов морфологического анализа в базовую архитектуру модели позволила достичь высоких показателей эффективности: точность 0.89, полнота 0.85, F1-мера 0.87, что подтверждает продуктивность комбинированного подхода к идентификации англицизмов в русскоязычных текстах.

Исследовательский проект М.В. Ивановой "Применение глубокого обучения для анализа заимствований" [25], реализованный в 2023 году, демонстрирует еще более впечатляющие результаты при комбинированном применении архитектуры RuRoBERTa-large и библиотеки SpaCy для анализа корпуса новостных текстов объемом 100000 предложений. Фокусирование исследовательского внимания на анализе контекстуального окружения лексических единиц позволило достичь исключительно высоких показателей эффективности: точность 0.92, полнота 0.88, F1-мера 0.90, что подтверждает критическую важность учета контекстуальных параметров при идентификации англицизмов в новостных текстах.

Сравнительный анализ различных методологических подходов к идентификации англицизмов, представленный в исследовательском проекте Г.О. Сидорова (2023) [33], предоставляет систематизированную информацию об эффективности различных алгоритмических решений в соответствующей области компьютерной лингвистики, что создает основу для формирования

оптимальной архитектуры аналитической системы с учетом специфики решаемых задач и особенностей анализируемого материала (таблица 3.1.3).

Таблица 3.1.3

Сравнительный анализ различных подходов к выявлению англицизмов

Метод	Точность	Полнота	F1-мера
Только правила	0.76	0.70	0.73
NLTK + правила	0.82	0.79	0.80
SpaCy + Word2Vec	0.85	0.83	0.84
RuBERT	0.89	0.87	0.88
RuRoBERTa + Natasha	0.93	0.91	0.92

Источник: Сидоров Г.О. Сравнительный анализ методов выявления англицизмов в текстах СМИ // Научно-технический вестник информационных технологий. 2023. Т. 23, № 1. С. 124-138.

Таким образом, англицизмы представляют собой многогранный лингвистический феномен, требующий комплексного подхода для автоматического анализа. Традиционные методы, основанные на правилах и словарях, демонстрируют ограниченную эффективность из-за динамической природы заимствований. Современные подходы, использующие алгоритмы машинного обучения и нейронные сети, особенно трансформерные модели, показывают значительно лучшие результаты. Анализ существующих инструментов выявил, что RuRoBERTa-large в сочетании с морфологическим анализом Natasha обеспечивает наивысшую точность при идентификации англицизмов в новостном дискурсе. На основании проведенного анализа представляется целесообразным разработать гибридную трехэтапную модель, где каждый этап оптимизирован для решения конкретной подзадачи в общем процессе обработки текста. Первый этап должен включать предпроцессинг с использованием специализированных библиотек и морфологический анализ, второй – эффективную классификацию с применением современных трансформерных архитектур, третий – генерацию адекватных русскоязычных эквивалентов с сохранением семантической целостности текста.

2. Разработка модели для автоматического выявления англицизмов в русскоязычных новостных статьях

2.1. Подготовка корпуса текстов для обучения и тестирования модели

Методологическая база исследования англицизмов в русскоязычных новостных публикациях характеризуется комплексным подходом к формированию репрезентативного текстового корпуса и разработкой многоступенчатой алгоритмической системы идентификации заимствованной лексики. Эмпирической основой исследования выступает масштабный датасет, сформированный на базе новостных материалов информационного агентства RBC за период с 1 апреля 2024 года по 1 апреля 2025 года [40], включающий 50793 текстовых единицы с подробной метаинформацией, охватывающей различные параметры публикаций: идентификационные данные, тематическую принадлежность, структурные характеристики, хронологические маркеры и гипертекстовые связи: ['id', 'project', 'project_nick', 'type', 'category', 'title', 'body', 'publish_date', 'publish_date_t', 'fronturl', 'overview', 'text']. Получили следующее распределение по категориям (таблица 2.1.1):

Таблица 2.1.1

Распределение по категориям

Ктегории	Доля, %
Политика	26310
Общество	11732
Спорт	6953
Бизнес	2017
Технологии и медиа	1316
Экономика	1277
Финансы	1084
База знаний	59
Авто	11
Свое дело	3
Недвижимость	1

Архитектура разработанной системы идентификации англицизмов базируется на принципах последовательной многоэтапной фильтрации с

использованием предварительно сформированных лексикографических ресурсов и включает несколько функциональных модулей, обеспечивающих комплексный анализ текстового материала. Предпроцессинг текстовой информации реализуется через сегментацию на предложения и извлечение лексических единиц с использованием алгоритмов токенизации, адаптированных к особенностям русскоязычного текста. Морфологическая нормализация осуществляется посредством лемматизации, приводящей различные словоформы к единой базовой форме, что критически важно для эффективного анализа текстов на русском языке с его развитой системой словоизменения. Лексический фильтр обеспечивает сопоставление извлеченных лексических единиц с базой известных англицизмов, а система контекстуальных исключений позволяет минимизировать количество ложноположительных результатов через фильтрацию потенциальных ошибочных идентификаций.

Лексикографическим фундаментом системы выступает специализированный словарь англицизмов, созданный на основе "Словаря англицизмов русского языка" А.И. Дьякова с последующей доработкой и расширением. Лексикографический ресурс, включающий 18784 лексические единицы, представляет собой репрезентативную выборку англоязычных заимствований, функционирующих в современном русском языке [38]. Статистический анализ словарного материала выявляет определенные закономерности в структуре англицизмов: наиболее частотным выступает сочетание начальной буквы "c" и финальной буквы "p" (рисунок 2.1.1), а длина лексических единиц преимущественно составляет 5 или 8-9 символов (рисунок 2.1.2), что отражает типологические особенности английских заимствований в русском языке.

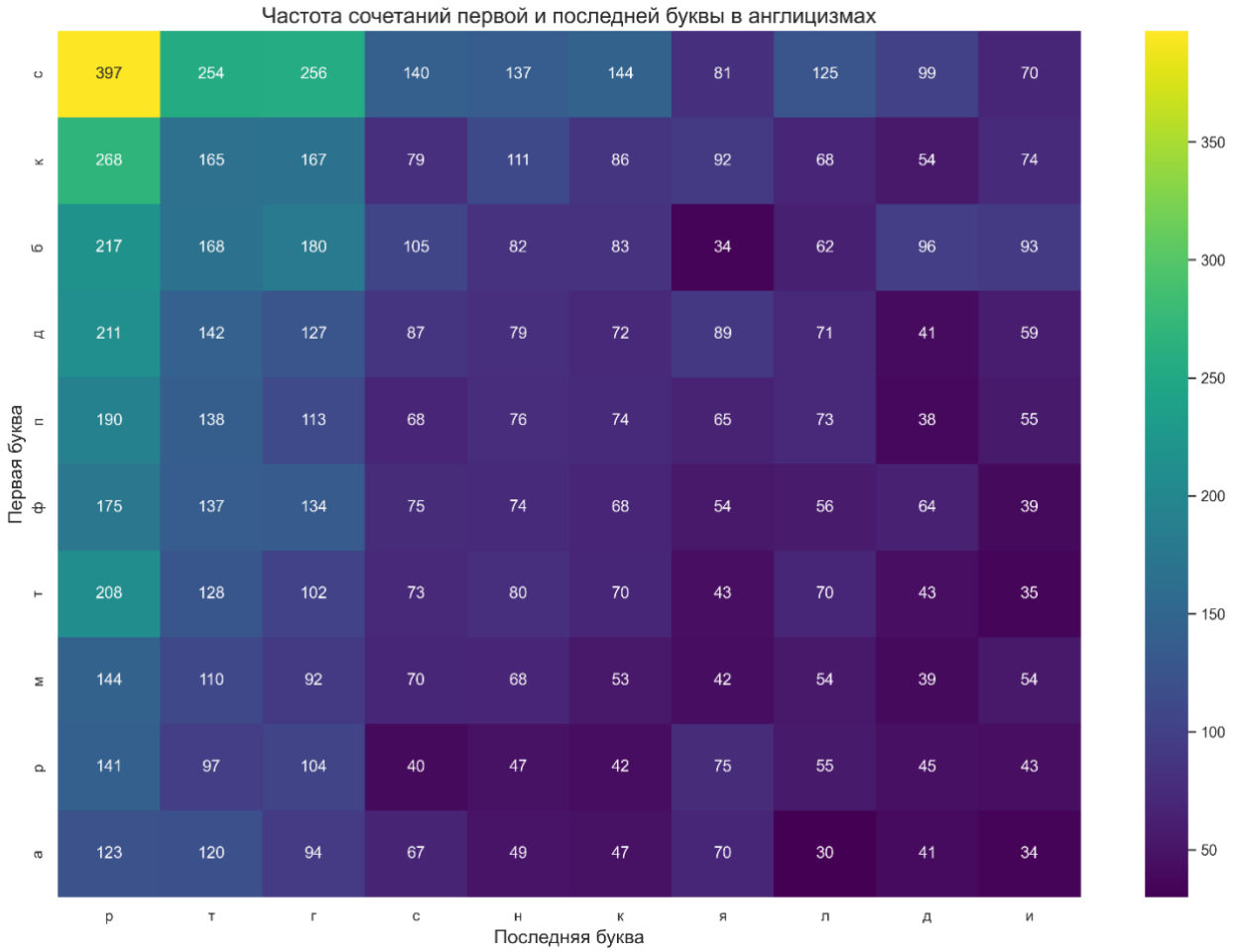


Рисунок 2.1.1. Частота сочетаний первой и последней буквы в англицизмах

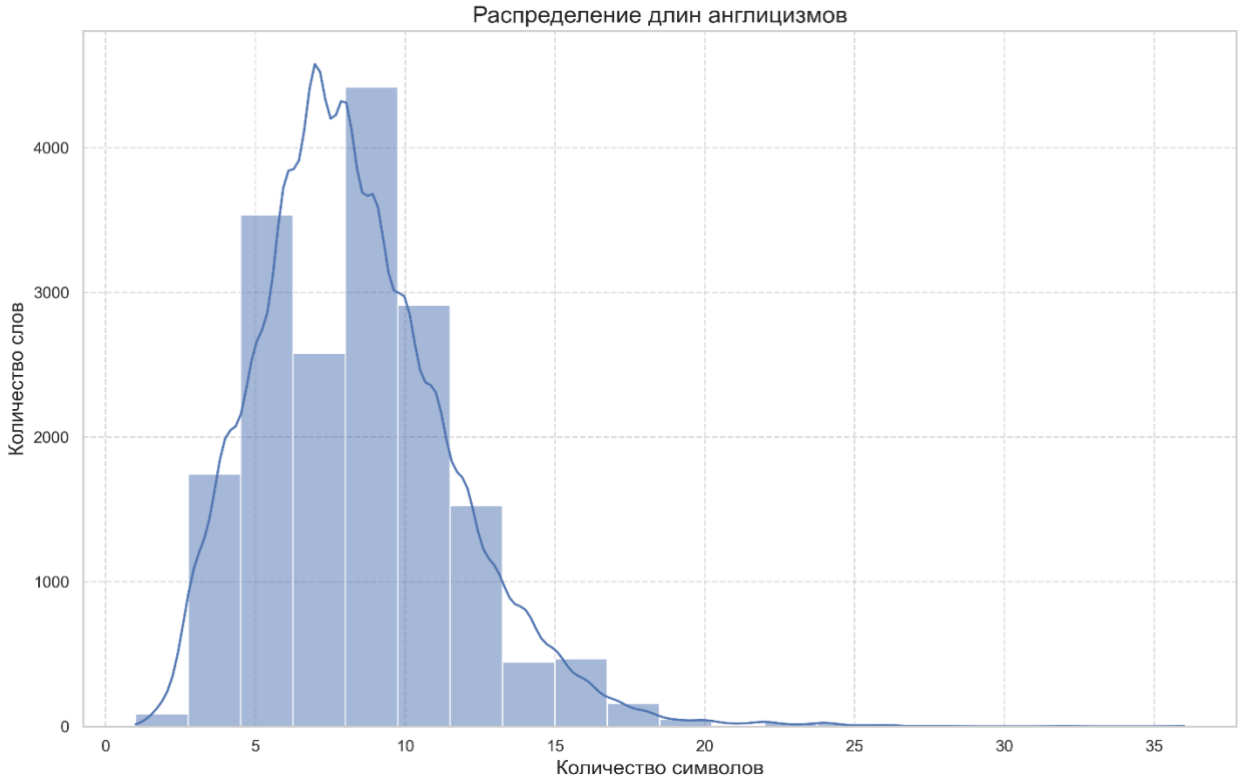


Рисунок 2.1.2. Распределение длин англицизмов

Данный словарь проходит дополнительную фильтрацию с исключением стоп-слов, что позволяет минимизировать шум при идентификации (рисунок 2.1.3):

```
all_anglicisms = [line.strip().lower() for line in f]
filtered_anglicisms = []
for word in all_anglicisms:
    word_lemma = lemmatize_word(word)
    if word_lemma not in stopwords_lemmas:
        filtered_anglicisms.append(word_lemma)
```

Рис. 2.1.3. Словарь англицизмов

Обработанный словарь представляет собой множество лемматизированных форм англицизмов, что обеспечивает эффективный поиск независимо от морфологических форм искомых слов в тексте.

Для решения задачи морфологической нормализации применяется лемматизация с использованием библиотеки Natasha, специализированной для обработки русскоязычных текстов и учитывающей особенности русской грамматики. Лемматизация существенно повышает полноту обнаружения англицизмов, приводя все формы слов к единой базовой форме, что особенно важно для русского языка с его развитой системой словоизменения и словообразования, позволяя идентифицировать англицизмы в различных грамматических формах и производных образованиях (рисунок 2.1.4):

```
def lemmatize_word(word):
    doc = Doc(word)
    doc.segment(segmenter)
    doc.tag_morph(morph_tagger)
    for token in doc.tokens:
        token.lemmatize(morph_vocab)
    if doc.tokens:
        return doc.tokens[0].lemma
    return word
```

Рис. 2.1.4. Лемматизация

Функциональным ядром системы выступает алгоритм find_anglicisms(), реализующий многоэтапный процесс анализа текстового материала:

1. Предварительно выполняется идентификация текста в различных типах кавычек («», "", ") для выявления собственных имен и цитат. Этот шаг позволяет избежать ложной идентификации имен собственных и брендов, которые часто пишутся в кавычках.

2. Для выделения слов используется регулярное выражение, учитывающее особенности русской орфографии, включая слова с дефисами.

3. Каждое найденное слово проходит многоуровневую проверку:

   A. Длина слова — игнорируются слова короче 3 символов:

   B. Имена собственные — если слово начинается с заглавной буквы и не стоит после точки, оно исключается из рассмотрения.

   C. Цифровые включения — слова, содержащие цифры, игнорируются.

После базовой фильтрации производится анализ на основе подготовленных словарей (рисунок 2.1.5):

```
if (word_lemma in anglicisms_base_forms and
        word_lemma not in stopwords_lemmas and
        word_lemma not in exceptions_lemmas and
        word_lemma not in quoted_lemmas):
    # Дополнительная проверка, не является ли слово частью текста в кавычках
    is_in_quotes = False
    for quoted_word in quoted_words:
        if word.lower() == quoted_word.lower():
            is_in_quotes = True
            break

    # Если слово не в кавычках и ещё не добавлено, добавляем его
    if not is_in_quotes and word not in found_anglicisms:
        found_anglicisms.append(word)
```

Рис. 2.1.5 Лексический анализ кандидатов

Каждая идентифицированная лексическая единица подвергается многоуровневой проверке с учетом различных параметров. Анализируется длина слова с исключением из рассмотрения лексических единиц короче трех символов, что позволяет фильтровать служебные слова и другие нерелевантные элементы. Учитывается графическое оформление слова с исключением из анализа имен собственных, начинающихся с заглавной буквы и не стоящих после точки, что позволяет избежать ложной идентификации антропонимов, топонимов и других проприальных единиц. Слова, содержащие цифровые включения, также игнорируются в процессе анализа, что повышает точность идентификации англицизмов в текстовом материале.

После базовой фильтрации осуществляется углубленный анализ на основе подготовленных лексикографических ресурсов с проверкой нескольких ключевых параметров. Верифицируется наличие леммы анализируемого слова в базе англицизмов, что позволяет идентифицировать заимствования на основе их нормализованных форм. Проверяется отсутствие лексической единицы в стоп-словах, исключающих из анализа функциональные слова и другие нерелевантные элементы. Контролируется отсутствие слова в списке исключений, содержащем лексические единицы,

ошибочно идентифицируемые как англицизмы из-за фонетического или орфографического сходства. Учитывается контекстуальное окружение слова с исключением из анализа лексических единиц, являющихся частью текста в кавычках, что позволяет избежать ложной идентификации цитат, названий и других специфических текстовых элементов.

После первичной идентификации англицизмов реализуется этап дополнительной фильтрации текстового материала для формирования финального датасета. Исключение дубликатов предложений обеспечивает повышение разнообразия корпуса и устранение повторяющихся элементов, что увеличивает репрезентативность текстовой выборки. Фильтрация предложений без англицизмов позволяет сформировать целевой датасет, содержащий текстовые фрагменты с идентифицированными заимствованиями, что оптимизирует процесс последующего анализа и интерпретации полученных данных.

Результатом применения разработанной методологии стало формирование специализированного набора данных, содержащего 50793 текстовые единицы с идентифицированными англицизмами, представленными в структурированном формате с текстом статьи в первом столбце и соответствующим англицизмом во втором, что обеспечивает удобство дальнейшего анализа и интерпретации полученных результатов в контексте исследования функционирования заимствованной лексики в русскоязычных новостных публикациях (рисунок 2.1.6).

	C1	C2
1	Поводов для эвакуации жителей Харьк…	["интервью"]
2	Там также есть интернет и аптечки.	["интернет"]
3	По словам мэра, за два дня последне…	["блэкаута"]
4	Утром 22 марта министр энергетики У…	["систем"]
5	Минобороны России тогда же отчитало…	["военно-промышленного", "комплекса…
6	В Белгороде и Белгородском районе д…	["системы"]
7	Венесуэльский YouTube-блогер Оскар …	["блогер", "роликами"]
8	Алехандро был арестован в городском…	["парк"]
9	На YouTube у блогера 1,89 млн подпи…	["блогера"]
10	The Hindustan Times сообщила, что а…	["аккаунты", "социальной"]

Рис. 2.1.6. Датасет из пар предложение-англицизмы

2.2. Разработка гибридной модели детекции и замены англицизмов: эксперименты на этапе классификации

Разработка оптимальной методологии идентификации англицизмов в русскоязычных новостных публикациях требует имплементации интегративного подхода, сочетающего различные алгоритмические решения в рамках многокомпонентной системы автоматического анализа текстовой информации. Комплексная архитектура предлагаемого решения включает несколько последовательных этапов обработки языкового материала, каждый из которых направлен на решение специфических задач в процессе выявления и замены англоязычных заимствований с сохранением смысловой целостности исходного текста. Многоаспектность поставленной задачи обуславливает необходимость синтеза различных технологий машинного обучения и компьютерной лингвистики, обеспечивающих эффективное решение отдельных компонентов общей проблемы идентификации заимствованной лексики.

Предлагаемая методологическая концепция базируется на трехэтапной архитектуре аналитической системы, включающей модули детекции англицизмов в текстовых массивах, генерации адекватных русскоязычных

эквивалентов и трансформации текстового материала с учетом грамматических особенностей русского языка. Селекция оптимального инструментария для каждого функционального модуля осуществлялась на основе многокритериального анализа с учетом таких параметров, как точность идентификации, вычислительная эффективность, ресурсные требования, интеграционный потенциал и применимость к русскоязычным текстовым массивам различной жанровой принадлежности.

Первый этап — детекция англицизмов в текстовом материале — представляет собой фундаментальную классификационную задачу, концептуально близкую к проблеме распознавания именованных сущностей (Named Entity Recognition, NER), требующую принятия бинарного решения относительно принадлежности каждой лексической единицы к категории англоязычных заимствований. Методологический арсенал для решения данной задачи включает два основных подхода: лексикографический метод, основанный на использовании предварительно составленных словарей, и алгоритмы машинного обучения, базирующиеся на статистическом анализе лингвистических характеристик текстового материала.

Лексикографический подход, несмотря на концептуальную простоту и вычислительную эффективность, демонстрирует существенные методологические ограничения при практическом применении. Экспериментальное тестирование показало, что точность данного метода составляет около 0.91, что заметно ниже показателей, достигаемых при использовании алгоритмов машинного обучения. Ключевым недостатком словарного подхода выступает низкий показатель полноты (recall) — около 0.65, означающий, что приблизительно 35% англицизмов остаются невыявленными в процессе анализа. Данное ограничение обусловлено принципиальной неспособностью лексикографического метода идентифицировать новейшие заимствования или их морфологические варианты, отсутствующие в предварительно составленном словаре, что в

условиях динамичного развития языка и регулярного появления новых заимствований представляется критическим методологическим недостатком.

Альтернативный подход, основанный на применении алгоритмов машинного обучения, включает такие методы, как логистическая регрессия, случайный лес (Random Forest) и градиентный бустинг, позволяющие идентифицировать англицизмы на основе анализа их морфологических, фонетических и контекстуальных характеристик. Для каждой лексической единицы формируется многомерный вектор признаков, включающий разнообразные параметры, характеризующие структурно-семантические особенности анализируемого слова и его контекстуального окружения.

Экспериментальная верификация эффективности различных методологических подходов осуществлялась на специально подготовленном сбалансированном датасете, включающем эквивалентное количество примеров англицизмов и неанглицизмов в соотношении 1:1. Для обучения и тестирования моделей машинного обучения была сформирована репрезентативная выборка из 20000 лексических единиц, из которых 18000 использовались для обучения алгоритмов и 2000 для тестирования их эффективности. Структура признакового пространства для моделей машинного обучения представлена в таблице 2.2.1, включающей различные параметры, характеризующие лингвистические особенности анализируемых лексических единиц.

Формирование векторных представлений лексических единиц и их контекстуального окружения осуществлялось с использованием трансформерной модели RuBERT, позволяющей получить контекстуализированные эмбеддинги размерностью 768, учитывающие семантико-синтаксические особенности анализируемого текстового материала. Дополнительно учитывались морфологические характеристики, извлеченные с помощью специализированной библиотеки Natasha, оптимизированной для работы с русскоязычными текстами, а также набор дополнительных признаков, включая частотные характеристики букв, наличие

сдвоенных буквенных комбинаций и особенности контекстуального окружения анализируемых лексических единиц. Общая размерность признакового пространства составила приблизительно 4799 параметров, что обеспечило комплексный учет морфологических и контекстуальных аспектов анализируемых лексических единиц при их классификации.

Категориальные признаки, такие как морфологические характеристики лексических единиц (падеж, род, число и другие грамматические категории), подвергались преобразованию с использованием техники One-Hot Encoding, трансформирующей категориальные переменные в систему бинарных признаков для их эффективной обработки алгоритмами машинного обучения. Для числовых признаков в случае логистической регрессии применялась процедура стандартизации данных, обеспечивающая повышение эффективности процесса обучения модели и улучшение конвергенции алгоритма оптимизации параметров.

Таблица 2.2.1

Структура признаков для моделей машинного обучения

Категория признаков	Описание	Количество признаков	Предобработка
Морфологические признаки			
Длина слова	Количество символов в слове	1	-
Регистр	Наличие заглавной буквы (is_capitalized)	1	-
Морфологические категории	Одушевленность (Animacy), падеж (Case), род (Gender), число (Number) и др.	~30	One-Hot Encoding
Контекстуальные признаки			
Соседние слова	Данные о двух словах слева и двух словах справа	4	-
Морфологические категории соседних слов	Морфологические признаки соседних слов (left_left_, *left_*, right_, *right_right_*)	~120	One-Hot Encoding
Длина соседних слов	Длина окружающих слов	4	-
Наличие англицизмов в контексте	Признаки is_anglicism для соседних слов	4	-
Символьные признаки			

Частота букв	Количество каждой буквы русского алфавита в слове	33	-
Сдвоенные буквы	Количество сдвоенных букв (аа, бб, вв, ...)	33	-
Векторные представления слов			
Эмбеддинги слова	Векторные представления слова из RuBERT (word_bert_*)	768	-
Эмбеддинги леммы	Векторные представления леммы из RuBERT (lemma_bert_*)	768	-
Эмбеддинги контекста	Векторные представления соседних слов из RuBERT	4 × 768 = 3072	-
Итого		~4799	

Анализ показателей вариационно-инфляционного фактора (VIF) выявил значительную мультиколлинеарность в структуре признакового пространства модели классификации англицизмов, что потенциально могло снизить эффективность алгоритмического решения. Признак "length", характеризующий длину лексической единицы и имеющий значение VIF > 10, продемонстрировал критически высокую линейную зависимость с другими переменными модели, что объясняется естественной корреляцией между длиной слова и частотными характеристиками входящих в него букв (count_о: 1.98, count_n: 2.80, count_а: 4.39, count_и: 4.44, count_к: 4.51, count_е: 5.33, count_с: 6.49) (рисунок 2.2.1). Увеличение количества символов в лексической единице закономерно приводит к возрастанию частоты встречаемости отдельных букв, что создает статистическую взаимозависимость между соответствующими признаками. Наблюдается также корреляция между длиной слова и количеством сдвоенных буквенных комбинаций, поскольку в более протяженных лексических единицах вероятность появления таких комбинаций объективно выше. Методологически обоснованным решением в данной ситуации выступает исключение признака длины слова из модели, поскольку данная характеристика фактически дублируется совокупностью частотных показателей отдельных букв, при этом последние обеспечивают

более детализированную информацию о морфологической структуре анализируемых лексических единиц.

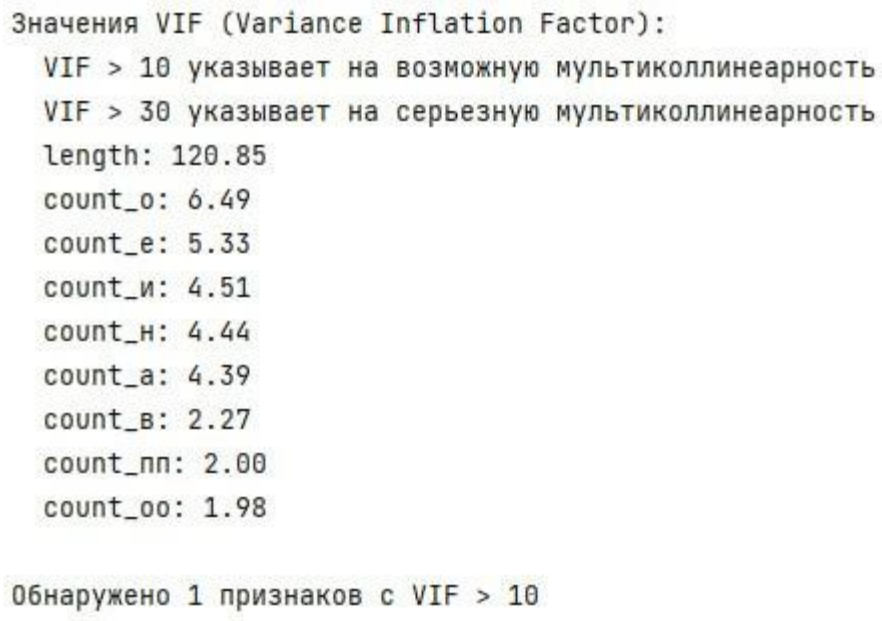


Рисунок 2.2.1. Результаты теста на мультиколлинеарность

Экспериментальное тестирование эффективности различных алгоритмических решений включало три основные модели: логистическую регрессию, градиентный бустинг и случайный лес, каждая из которых характеризуется специфическими особенностями функционирования и областями эффективного применения.

Логистическая регрессия представляет собой линейную классификационную модель, в которой вероятность принадлежности лексической единицы к категории англицизмов рассчитывается согласно формуле:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}, (2.2.1)$$

где Y — бинарная переменная, принимающая значение а для англицизмов и 0 для неанглицизмов, P(Y=1|X) — вероятность принадлежности слова к категории англицизмов при заданном наборе признаков X, β_0 — свободный член модели, β — вектор коэффициентов,

определяемых в процессе обучения путем максимизации функции правдоподобия:

$$L(\beta) = \prod_{i=1}^{\text{ь}} P\left(y^{(i)} \mid x^{i}; \beta\right)^{y^{(i)}} \left(1 - P\left(y^{(i)} \mid x^{(i)}; \beta\right)\right)^{1-y(i)}, (2.2.2)$$

где m – количество слов в обучающей выборке.

К преимуществам логистической регрессии относятся высокая интерпретируемость результатов, вычислительная эффективность и относительно низкие требования к объему обучающих данных. Анализ коэффициентов модели позволяет выявить признаки, наиболее значимые для идентификации англицизмов, что имеет существенное значение для понимания лингвистических механизмов заимствования и разработки более эффективных алгоритмических решений.

Экспериментальное тестирование показало, что логистическая регрессия демонстрирует более высокие результаты по сравнению с лексикографическим методом, достигая точности 0.9465 и полноты 0.9530, что свидетельствует о существенном превосходстве данного алгоритмического решения над традиционным словарным подходом. Линейная природа модели обеспечивает высокую интерпретируемость результатов, позволяя выявить наиболее значимые признаки для классификации англицизмов и понять лингвистические механизмы их функционирования в русскоязычных текстах.

Анализ коэффициентов логистической регрессии выявил ряд интересных лингвистических закономерностей. Наличие заглавной буквы в начале слова выступает сильным отрицательным предиктором (коэффициент -2.304354), что логически объяснимо, поскольку заглавная буква часто маркирует имена собственные или начало предложения, снижая вероятность принадлежности слова к категории англицизмов. Значимыми предикторами также выступают определенные компоненты векторных представлений из модели RuBERT и частотные характеристики редких букв русского алфавита, таких как "щ" (коэффициент -1.088343) и "ф" (коэффициент 0.936681).

Последнее наблюдение соответствует лингвистическим особенностям русского языка, где буква "ф" часто встречается в заимствованных словах, что создает статистически значимую корреляцию между данным признаком и принадлежностью лексической единицы к категории англицизмов.

ROC-кривая (Receiver Operating Characteristic curve) модели логистической регрессии, представляющая собой графическое отображение эффективности бинарного классификатора при различных пороговых значениях, имеет выраженную форму, приближающуюся к идеальной конфигурации (рисунок 2.2.2). Площадь под кривой (AUC, Area Under Curve) составляет 0.982, что является исключительно высоким показателем, близким к теоретическому максимуму 1.0, и свидетельствует о высокой дискриминационной способности модели при идентификации англицизмов в русскоязычных текстах.

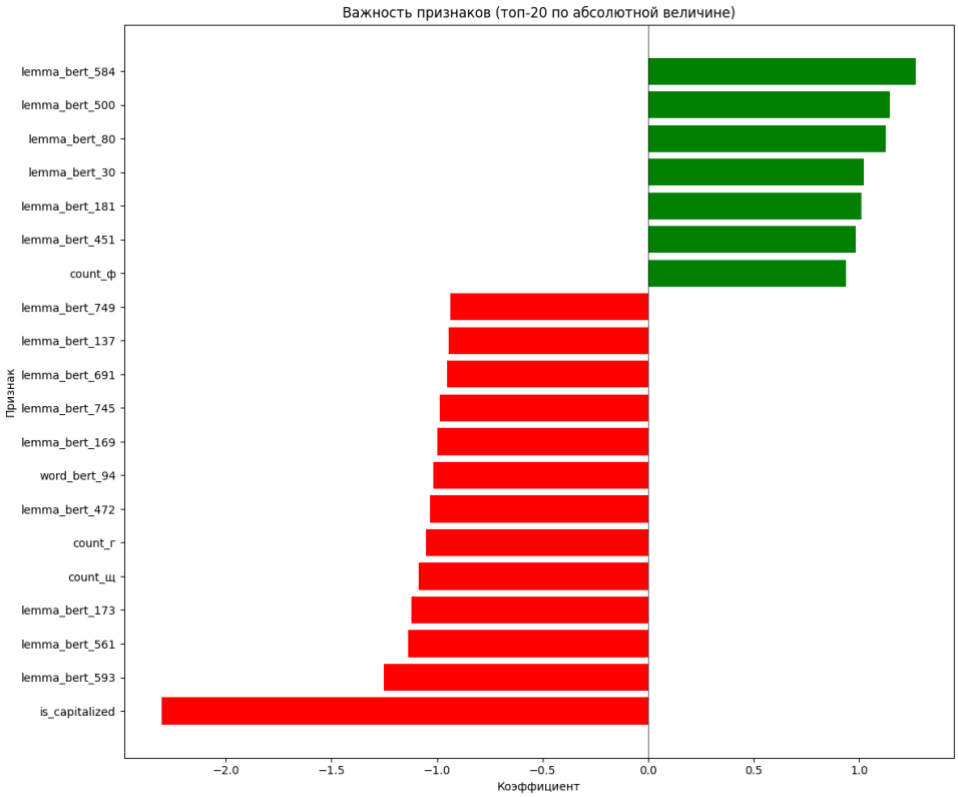


Рисунок 2.2.2. Важность признаков Logistic Regression

Сравнительный анализ различных алгоритмических подходов к задаче идентификации англицизмов в русскоязычных текстах демонстрирует

значительные различия в эффективности, вычислительных затратах и интерпретируемости результатов, что позволяет определить оптимальные методологические решения для создания высокоточной системы автоматического анализа заимствованной лексики.

Алгоритм Random Forest, относящийся к категории ансамблевых методов машинного обучения, базируется на принципе построения множества независимых деревьев решений с последующей агрегацией их прогностических результатов для формирования финального решения. Каждое дерево в ансамбле обучается на бутстрэп-выборке из исходного набора данных, а при формировании узлов рассматривается случайное подмножество признаков, что обеспечивает диверсификацию индивидуальных классификаторов в ансамбле и повышает его обобщающую способность. Вероятность принадлежности лексической единицы к категории англицизмов определяется как доля деревьев в ансамбле, классифицировавших ее соответствующим образом:

$$\hat{\mathrm{f}} = \frac{1}{\mathrm{B}} \sum_{b=1}^{B} f_b(x)\,, (2.2.3)$$

где $\hat{\mathrm{f}}$ – итоговое предсказание, B – количество деревьев в ансамбле, $f_b(x)$ – предсказание b-го дерева для входного вектора признаков x ((1 – англицизм, 0 – не англицизм).

Ключевыми преимуществами алгоритма Random Forest выступают высокая устойчивость к переобучению, эффективное моделирование нелинейных зависимостей между признаками и стабильная работа в условиях разреженных данных, характерных для задач обработки естественного языка. Алгоритм позволяет оценивать значимость различных признаков для классификации через методологии среднего уменьшения нечистоты узлов (Mean Decrease in Impurity, MDI) или среднего уменьшения точности (Mean Decrease in Accuracy, MDA), что обеспечивает определенный уровень интерпретируемости результатов.

Анализ зависимости точности алгоритма Random Forest от количества деревьев в ансамбле показал, что оптимальная конфигурация модели включает 200-250 базовых классификаторов, после чего наблюдается эффект переобучения (рисунок 2.2.3), снижающий обобщающую способность модели. Данная закономерность типична для ансамблевых методов, где увеличение количества составных моделей не всегда приводит к пропорциональному повышению эффективности классификации.

Рисунок 2.2.3. Зависимость точности от деревьев

Градиентный бустинг (Gradient Boosting) представляет собой еще один ансамблевый метод машинного обучения, принципиально отличающийся от Random Forest последовательным, а не параллельным формированием составных моделей, где каждый новый классификатор оптимизируется для минимизации ошибок, допущенных предыдущими компонентами ансамбля. Математическая формализация данного подхода может быть представлена как

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x), (2.4)$$

где F_m(x) — модель после m итераций, F_{m-1}(x) — модель после m−1 итераций, η — скорость обучения (learning rate), h_m(x) — слабый

классификатор (обычно дерево решений), обученный на остаточных ошибках предыдущей модели.

Процесс обучения градиентного бустинга основан на минимизации функции потерь L путем последовательного добавления моделей, следующих антиградиенту данной функции:

$$h_m = arg\ min_h \sum_{i=1}^{n} [-g_i * h(x_i)], (2.5)$$

где $g_i = \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}, (2.6)$ – градиент функции потерь для i-го наблюдения.

Экспериментальное тестирование продемонстрировало, что реализация градиентного бустинга в виде алгоритма XGBoost показывает наивысшие результаты среди всех исследованных методов, достигая впечатляющих показателей точности (0.9885), точности в терминах precision (0.9929) и полноты (0.9840), что в совокупности обеспечивает исключительно высокое значение интегрального F1-критерия (0.9884). Такие выдающиеся результаты объясняются способностью XGBoost эффективно моделировать комплексные нелинейные зависимости между признаками и постепенно фокусироваться на сложных для классификации примерах, что особенно важно в контексте идентификации англицизмов, характеризующихся значительной вариативностью форм и контекстуального употребления. Временные затраты на обучение XGBoost составляют приблизительно 2-4 минуты, что делает его более эффективным по сравнению с Random Forest, несмотря на последовательную природу алгоритма, ограничивающую возможности параллелизации вычислений (таблица 2.2.2).

Таблица 2.2.2

Сравнение метрик эффективности модели на экспериментальном наборе данных

Метрика	Логистическая регрессия	Random Forest	XGBoost

Точность (Accuracy)	0.9465	0.9775	0.9885
Точность (Precision)	0.9408	0.9948	0.9929
Полнота (Recall)	0.9530	0.9600	0.9840
F1-мера	0.9468	0.9771	0.9884

Анализ значимости признаков в модели XGBoost (рисунок 2.2.4, 2.2.5) демонстрирует более равномерное распределение по сравнению с Random Forest, что свидетельствует о более комплексном использовании информации из различных источников и эффективном моделировании взаимодействий между признаками. Наиболее значимыми предикторами выступают компоненты векторных представлений слов, извлеченные из трансформерной модели RuBERT, что подтверждает эффективность использования предобученных языковых моделей для извлечения семантически значимых характеристик лексических единиц и их контекстуального окружения.

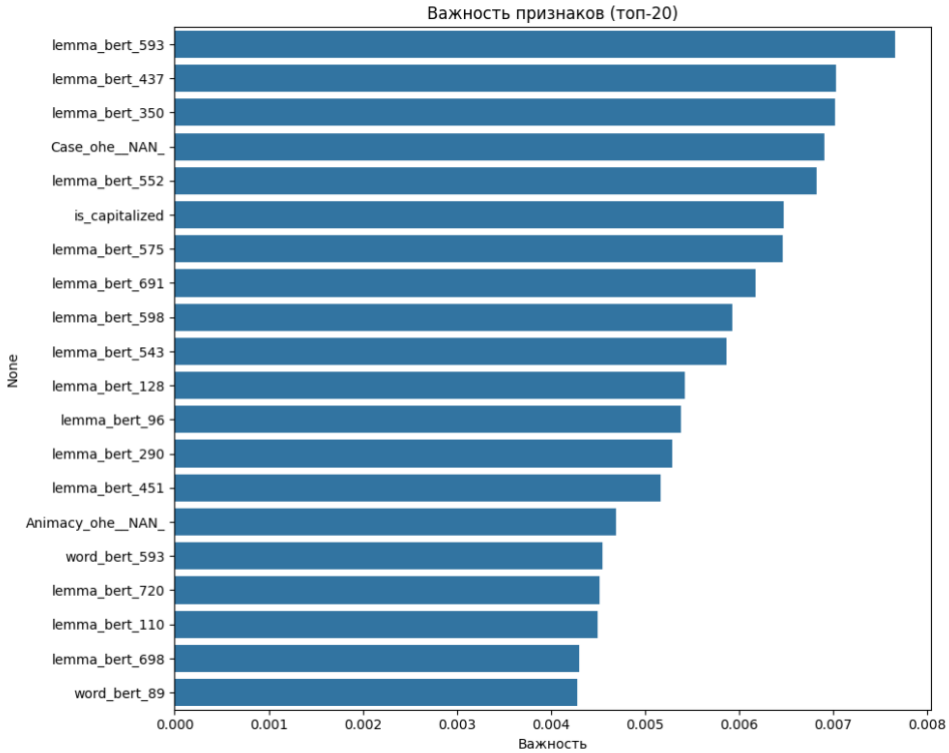


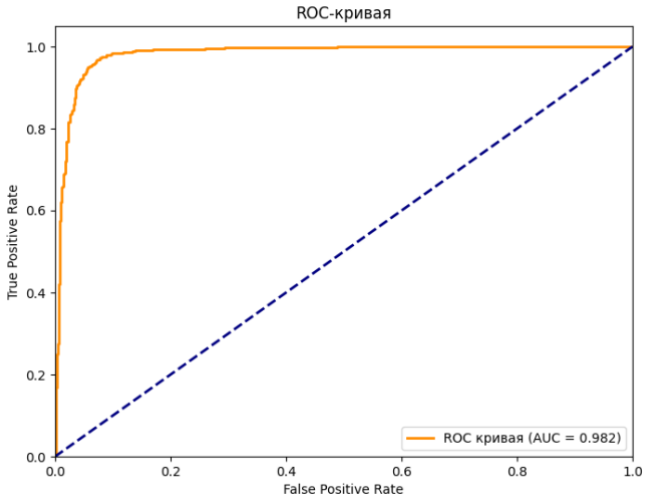
Рисунок 2.2.4. Важность признаков RF

Рисунок 2.2.5. Важность признаков XgBoost

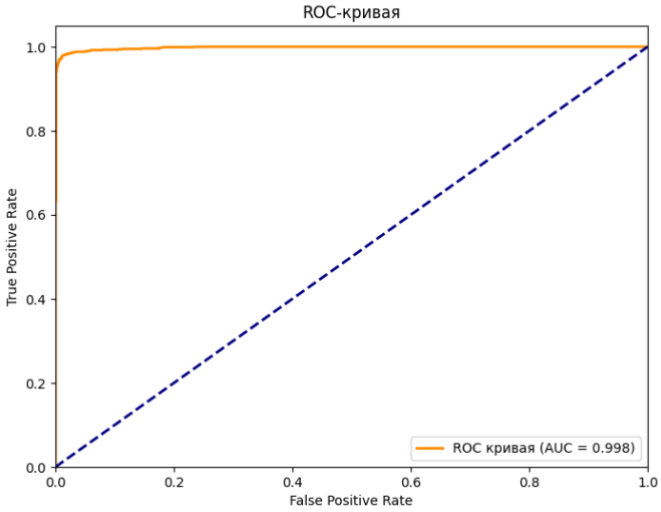
Random Forest демонстрирует значительное превосходство над логистической регрессией по основным метрикам эффективности, достигая точности 0.9775 и исключительно высокого значения precision (0.9948), что означает практически полное отсутствие ложноположительных результатов — случаев, когда модель ошибочно классифицирует неанглицизмы как заимствования. Однако показатель полноты (recall) алгоритма (0.9600) несколько ниже, чем у XGBoost, что указывает на пропуск приблизительно 4% англицизмов в процессе анализа текстового материала и создает определенные ограничения для применения данного метода в системах, где критически важно выявление максимального количества заимствований.

Сравнительный анализ динамики обучения для всех трех алгоритмических подходов свидетельствует о значительном преимуществе ансамблевых методов над логистической регрессией уже на начальных этапах обучения, при этом XGBoost демонстрирует наиболее стабильный рост

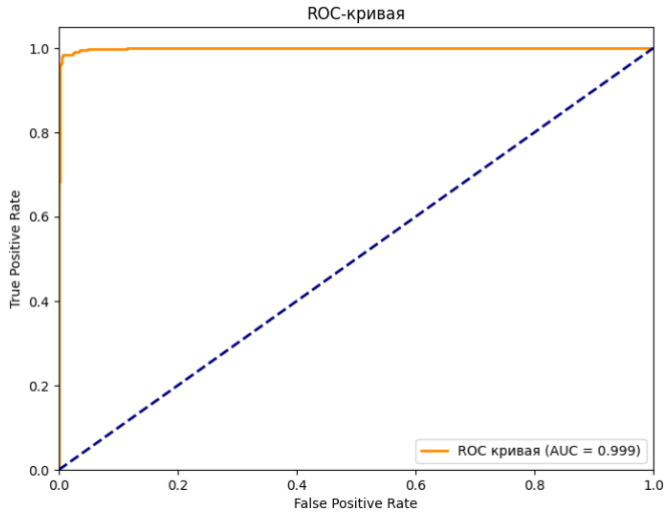
эффективности с увеличением объема обучающих данных, что указывает на высокую обобщающую способность алгоритма при работе с новыми, ранее не встречавшимися примерами, и делает его оптимальным выбором для практической реализации системы идентификации англицизмов в русскоязычных текстах (рисунок 2.2.6).



а) Logistic Regression



б) XgBoost



б) Random Forest

Рисунок 2.2.6 ROC-кривая RF

Для комплексной оценки эффективности рассмотренных методологических подходов была проведена серия экспериментов на тестовом наборе данных, содержащем русскоязычные тексты с англицизмами различных категорий. Результаты экспериментального тестирования позволили определить оптимальную комбинацию методов для решения задачи идентификации заимствований в текстовых массивах. В качестве критериев сравнения использовались такие метрики, как точность (precision), полнота (recall), F1-мера, вычислительная эффективность, интерпретируемость результатов и устойчивость к разреженным данным, что обеспечило комплексную оценку различных аспектов функционирования алгоритмических решений в контексте поставленной задачи. Визуализация сравнительного анализа трех основных моделей представлена в таблице 2.2.3.

Таблица 2.2.3.

Сравнительный анализ методов классификации для детекции англицизмов

Критерий	Логистическая регрессия	Random Forest	Градиентный бустинг
Тип задачи	Классификация токенов (is-anglicism: 0/1)	Классификация токенов (is-anglicism: 0/1)	Классификация токенов (is-anglicism: 0/1)

Сложность подготовки данных	Средняя (требуется предварительная обработка текста, векторизация признаков)	Средняя (требуется предварительная обработка текста, векторизация признаков)	Средняя (требуется предварительная обработка текста, векторизация признаков)
Интерпретируемость	Высокая (коэффициенты модели имеют ясную интерпретацию)	Средняя (важность признаков может быть оценена, но интерпретация сложнее)	Средняя (важность признаков может быть оценена, но интерпретация сложнее)
Вычислительная эффективность	Высокая (быстрое обучение и предсказание)	Средняя (обучение может быть ресурсоемким при большом числе деревьев)	Низкая (последовательное обучение требует значительных ресурсов)
Способность к моделированию нелинейных зависимостей	Низкая (только линейные границы)	Высокая (эффективное моделирование сложных взаимодействий)	Высокая (последовательное улучшение предсказаний)
Устойчивость к переобучению	Средняя (требуется регуляризация)	Высокая (встроенная регуляризация через усреднение)	Средняя (требуется тщательная настройка параметров)
Обработка несбалансированных данных	Слабая (требуются дополнительные техники)	Средняя (можно настроить через весовые коэффициенты)	Сильная (встроенные механизмы взвешивания)
Работа с разреженными данными	Эффективная	Эффективная	Эффективная
Требования к объему обучающих данных	Низкие	Средние	Высокие
Параллелизация	Поддерживается для отдельных частей	Высокая (независимое обучение деревьев)	Ограниченная (последовательное обучение)
Скорость предсказания	Очень высокая	Высокая	Средняя
Требования к памяти	Низкие	Высокие (хранение всех деревьев)	Средние

Архитектура алгоритмов машинного обучения для выявления англицизмов может быть представлена следующим образом (рисунок 2.2.7):

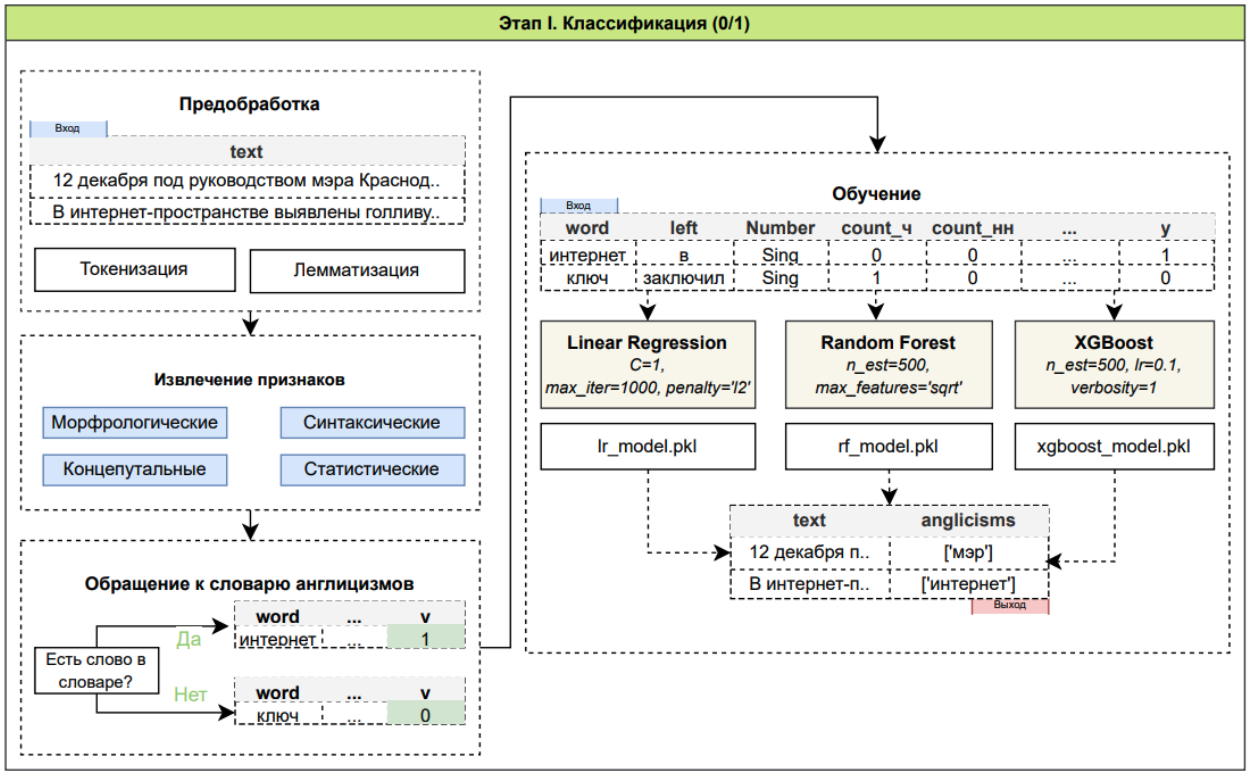


Рисунок 2.2.8. Архитектура моделей машинного обучения для детекции англицизмов

Левая часть архитектурной схемы отображает процесс предварительной обработки исходных текстовых данных, включающий этапы токенизации и лемматизации, что соответствует стандартным практикам предпроцессинга в задачах обработки естественного языка.

Правая часть архитектурной схемы отображает процесс обучения моделей машинного обучения, включающий три основных алгоритмических подхода: логистическую регрессию (Linear Regression), случайный лес (Random Forest) и XGBoost, что соответствует рассмотренным в исследовании методам классификации. Подобная архитектурная организация обеспечивает интеграцию различных методологических подходов в единую систему автоматического анализа англицизмов в русскоязычных текстах с возможностью выбора оптимального алгоритма в зависимости от специфики решаемой задачи и доступных вычислительных ресурсов.

2.3 Разработка гибридной модели детекции и замены англицизмов: эксперименты на этапе синонимирования

Второй этап методологического комплекса по идентификации и замене англицизмов в русскоязычных текстах сфокусирован на формировании эталонных текстовых образцов с заменой выявленных заимствований на семантически эквивалентные русскоязычные аналоги при сохранении грамматической и стилистической целостности исходного материала. Данная фаза исследования представляет собой критически важный компонент общей архитектуры системы, обеспечивающий возможность обучения модели, способной не только обнаруживать англоязычные заимствования, но и осуществлять их адекватную замену с минимальными смысловыми потерями и сохранением языковой естественности текста.

В качестве базовой архитектуры для решения данной задачи была выбрана современная языковая модель Qwen, разработанная корпорацией Alibaba, представляющая собой многоязычный трансформер с механизмом самовнимания, адаптированный для работы с разнообразными лингвистическими задачами. Математический формализм механизма самовнимания, лежащий в основе архитектуры трансформера, для каждого токена входной последовательности вычисляет взвешенную сумму всех токенов с весовыми коэффициентами, определяемыми скалярным произведением векторов запроса (query) и ключа (key):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{\text{d}_\text{k}}}\right)\text{V}, \quad (2.6)$$

где Q, K и V – матрицы запросов, ключей и значений соответственно; d_k – размерность векторов ключей.

Для количественной оценки семантического сходства между исходным предложением и его модифицированной версией с заменой англицизмов применяется специализированная модель SentenceTransformer, обученная на русскоязычном корпусе текстов для извлечения семантически значимых

векторных представлений текстовых фрагментов. Семантическая близость вычисляется как косинусная мера между векторными представлениями (эмбеддингами) анализируемых предложений:

$$similarity(s_1, s_2) = \frac{E(s_1) \cdot E(s_2)}{||E(s_1)|| \cdot ||E(s_2)||}, (2.8)$$

где E(s) – векторное представление предложения s, полученное с помощью энкодера модели SentenceTransformer. Данный подход обеспечивает объективную количественную оценку степени сохранения семантического содержания при замене англицизмов на русскоязычные эквиваленты, что позволяет выбирать оптимальные варианты замены с минимальными смысловыми искажениями.

Из линейки моделей QWEN для реализации системы была выбрана модификация Instruct, оптимизированная для точного выполнения инструкций, что особенно важно для моделей с ограниченными вычислительными параметрами, обеспечивая более детерминированное и контролируемое поведение при генерации русскоязычных эквивалентов англицизмов в соответствии с заданными критериями семантической и грамматической корректности.

Для обеспечения грамматической корректности заменяемых лексических единиц применяется дополнительный этап трансформации, учитывающий морфологические особенности русского языка и направленный на согласование заменяющего слова с его контекстуальным окружением. Данный этап реализуется с использованием модели Qwen с пониженным значением параметра temperature (0.3), что обеспечивает более детерминированную генерацию, ориентированную на грамматическую правильность за счет снижения вариативности выходных результатов. В результате применения описанной методологии для каждого предложения, содержащего англицизмы, формируется эталонный вариант с их заменой на русскоязычные эквиваленты, сохраняющий смысловое содержание, стилистические особенности и грамматическую структуру исходного текста.

Для определения оптимального баланса между вычислительной эффективностью и качеством генерации русскоязычных эквивалентов было проведено тестирование различных версий модели Qwen с вариативным количеством параметров: 0.5B, 1.5B, 3B, 7B [45, 46, 47]. Экспериментальные результаты демонстрируют существенные различия в производительности и качестве генерации для моделей различного масштаба. Особого внимания заслуживает наблюдение, что модели с меньшим количеством параметров (0.5B и 1.5B) демонстрируют недостаточное качество генерации русскоязычных эквивалентов, часто продуцируя альтернативные англицизмы или модификации исходного заимствования вместо адекватных русскоязычных синонимов (таблица 2.3.1). Данная особенность приводит к необходимости повторной генерации для достижения требуемого количества валидных русскоязычных эквивалентов, что существенно увеличивает среднее время обработки запроса и снижает эффективность системы в целом, несмотря на потенциально более высокую вычислительную производительность моделей с меньшим количеством параметров.

Таблица 2.3.1.

Сравнение моделей QWEN

Количество параметров	Среднее семантическое сходство, %	Среднее время на обработку одного предложения, секунд
0.5B	92.4	14
1.5B	94.8	11
3B	97.8	12
7B	98.2	38
14B	98.3	124

Экспериментальное тестирование различных конфигураций модели Qwen с вариативным количеством параметров позволило определить оптимальное соотношение между семантической точностью замены англицизмов и вычислительной эффективностью процесса обработки текстового материала. Наилучший баланс между данными параметрами

достигается при использовании модели с 3 миллиардами параметров, обеспечивающей высокий уровень семантического сходства между исходным и модифицированным предложениями (97.8%) при приемлемых временных затратах на обработку (12 секунд на предложение), что свидетельствует о практической применимости данной конфигурации в реальных системах анализа и обработки текстовой информации.

Модели с меньшим количеством параметров (0.5B и 1.5B) продемонстрировали парадоксально низкую эффективность как в аспекте семантической точности замены англицизмов, так и в отношении вычислительной производительности, несмотря на потенциально более высокую скорость обработки данных за счет сниженной параметрической сложности. Данный эффект объясняется недостаточным качеством генерируемых синонимов, что приводит к продуцированию альтернативных англицизмов или вариаций исходного заимствования вместо адекватных русскоязычных эквивалентов, что, в свою очередь, требует повторной генерации для достижения требуемого количества валидных замен и существенно увеличивает фактическое время обработки запроса.

Значительное увеличение параметрического объема модели до 7B и 14B сопровождается лишь незначительным повышением семантической точности замены англицизмов (до 98.2% и 98.3% соответственно) при экспоненциальном росте вычислительных затрат (38 и 124 секунды на предложение), что свидетельствует о существенном снижении предельной полезности увеличения размера модели и неэффективности использования более сложных конфигураций в практических сценариях обработки текстовой информации. Данное наблюдение подтверждает оптимальность выбора модели с 3 миллиардами параметров в качестве базовой архитектуры для системы замены англицизмов, обеспечивающей наилучшее соотношение между семантической точностью и вычислительной эффективностью.

2.4 Разработка гибридной модели детекции и замены англицизмов: эксперименты на этапе тонкой настройки языковой модели

Третий, заключительный этап методологического комплекса по замене англицизмов в русскоязычных текстах сфокусирован на обучении языковой модели с использованием сформированного корпуса эталонных пар предложений, включающих оригинальные тексты с англицизмами и их семантически эквивалентные варианты с русскоязычными заменами. Данный подход нацелен на создание полнофункциональной автоматизированной системы, способной обрабатывать произвольные текстовые массивы с минимальным участием человека. Основная задача данного этапа заключается в тонкой настройке крупной языковой модели (Large Language Model, LLM) на специфической задаче замещения англоязычных заимствований с комплексным учетом контекстуальных параметров, грамматических особенностей и стилистического единства обрабатываемого текстового материала.

Первичный этап реализации данной фазы исследования включал детальный анализ существующих языковых моделей для определения оптимальной архитектуры, соответствующей ряду критических требований: высокое качество обработки русскоязычных текстов, обеспечивающее адекватную работу с материалом целевого языка; достаточное понимание семантических нюансов для корректной замены англицизмов с сохранением смысловой целостности текста; приемлемые вычислительные требования, обеспечивающие возможность тонкой настройки на доступном оборудовании без необходимости использования специализированных высокопроизводительных вычислительных кластеров; потенциал для инкрементального обучения при расширении корпуса данных, что обеспечивает масштабируемость и адаптивность системы при появлении новых англицизмов и контекстов их использования.

На основе сформулированных критериев для сравнительного анализа были отобраны несколько репрезентативных моделей различной архитектурной сложности и функциональной специализации: Qwen2.5 (разработка Alibaba Cloud) в нескольких версиях с различным количеством параметров (1.5B, 3B, 7B), YandexGPT-5-Lite (разработка компании Яндекс) в версии 8B, TinyLlama (проект TinyLlama) в версии 1.1B, а также rugpt3large_based_on_gpt2 (разработка AI-Forever). Селекция данных моделей обусловлена необходимостью сравнительной оценки эффективности архитектурных решений различной параметрической сложности и функциональной специализации для решения специфической задачи идентификации и замены англицизмов в русскоязычных текстах.

Модель ai-forever/rugpt3large_based_on_gpt2 [42] представляет собой русскоязычную адаптацию архитектуры GPT-2, прошедшую специализированное обучение на масштабном корпусе русскоязычных текстов для оптимизации работы с материалом целевого языка. С архитектурной точки зрения данная модель базируется на механизме декодер-ориентированных трансформеров, содержит приблизительно 760 миллионов параметров и характеризуется сложной структурой с 24 слоями, 16 головами внимания и размерностью скрытого состояния 1280. Обучение модели проводилось на разнообразном корпусе русскоязычных интернет-текстов, собранных компанией SberAI, включающем материалы различных источников: Лента.ру, русскоязычный сегмент Википедии, RuBooks, открытые форумы, блоги и новостные сайты. Общий объем обучающего корпуса составляет порядка 600 ГБ текстового материала, что обеспечивает широкий охват языковых особенностей и контекстов использования англицизмов в русскоязычной среде.

Модель TinyLlama/TinyLlama-1.1B-Chat-v1.0 [51] представляет компактную реализацию архитектуры Llama, специально разработанную для обеспечения эффективной работы на устройствах с ограниченными вычислительными ресурсами при сохранении достаточного уровня качества

генерируемого текста. TinyLlama представляет собой компактную языковую модель, базирующуюся на архитектуре decoder-only Transformer, обеспечивающей совместимость с архитектурными решениями LLaMA (Large Language Model Meta AI). Техническая реализация модели включает использование Rotary Position Embeddings (RoPE) для эффективного кодирования позиционной информации, SwiGLU-активации в слоях прямого распространения, а также механизма multi-query attention для оптимизации процесса обработки информации. Основная цель данного проекта заключается в максимальной компрессии параметрической сложности модели до приблизительно 1.1 миллиарда параметров при минимизации потери качества генерируемого текстового материала и сохранении архитектурной совместимости с базовой моделью LLaMA, что обеспечивает возможность использования существующих методологических подходов и инструментария для работы с данной архитектурой.

Модель Qwen/Qwen2.5-1.5B-Instruct [45] относится к семейству инструктивно-настраиваемых языковых моделей, специфически оптимизированных для точного выполнения задач в соответствии с предоставленными инструкциями, что особенно важно для контролируемой замены англицизмов с учетом заданных критериев и ограничений. Qwen 2.5 представляет собой вторую версию семейства языковых моделей, разработанных Alibaba DAMO Academy, построенную на основе декодер-ориентированной трансформерной архитектуры, концептуально близкой к моделям семейства GPT. В версии модели с 1.5 миллиардами параметров используется 24 слоя трансформеров, 2048 скрытых единиц, 16 механизмов внимания, RMSNorm вместо традиционной LayerNorm для нормализации, Rotary Position Embeddings (RoPE) для кодирования позиционной информации, а также поддержка механизма multi-query attention для оптимизации обработки запросов.

Обучение базовой модели Qwen2.5 производилось на разнообразных многоязычных датасетах, включающих программный код, академические

тексты, веб-контент, материалы Википедии, разговорные данные, а также синтетически сгенерированные текстовые корпуса для расширения охвата лингвистических явлений. Для инструктивной версии модели (Instruct) дополнительно использовались специализированные пары "инструкция-результат", включающие задачи перевода, обобщения информации, вопросно-ответные системы и задачи программирования, что расширяет функциональные возможности модели и повышает ее эффективность при решении специфических задач, таких как замена англицизмов в русскоязычных текстах.

По сравнению с более параметрически сложными моделями семейства Qwen2.5, версия 1.5B-Instruct демонстрирует существенные преимущества с точки зрения вычислительной эффективности и доступности применения в условиях ограниченных ресурсов:

- Модель Qwen/Qwen2.5-3B-Instruct характеризуется удвоенным количеством параметров (3 миллиарда) и, соответственно, требует, как минимум вдвое больше видеопамяти при генерации текста и тонкой настройке, а при полноценном обучении — более агрессивных техник градиентного накопления, что существенно усложняет использование данной модели на доступных исследовательских платформах, таких как Colab с графическими процессорами T4.
- Модель Qwen/Qwen2.5-7B-Instruct представляет собой значительно более ресурсоемкую архитектуру с 7 миллиардами параметров, ориентированную на многокарточные конфигурации или серверные среды с профессиональными графическими процессорами типа A100 или V100 с объемом видеопамяти 40–80 ГБ.

Дополнительные экспериментальные тестирования модели YandexGPT-5-Lite-8B [53] показали, что, несмотря на заявленную оптимизацию, данная архитектура с 8 миллиардами параметров также не смогла эффективно

функционировать в стандартной среде Colab из-за критического недостатка видеопамяти (VRAM) и ограничений по времени выполнения операций, что создает существенные препятствия для ее использования в исследовательских целях без доступа к специализированной вычислительной инфраструктуре.

Проведенный анализ различных архитектур подчеркивает, что для исследовательских и прототипирующих целей в условиях ограниченных вычислительных ресурсов, таких как облачные платформы Colab и Kaggle, выбор модели Qwen2.5-1.5B-Instruct представляется не только компромиссным, но и практичным решением, особенно в контексте необходимости соблюдения баланса между качеством генерируемого текста и доступностью вычислительных ресурсов для эффективной реализации процесса обучения и применения модели.

Отобразим основные этапы реализации экспериментов по генерации предложений без англицизмов на рисунке 2.4.1.

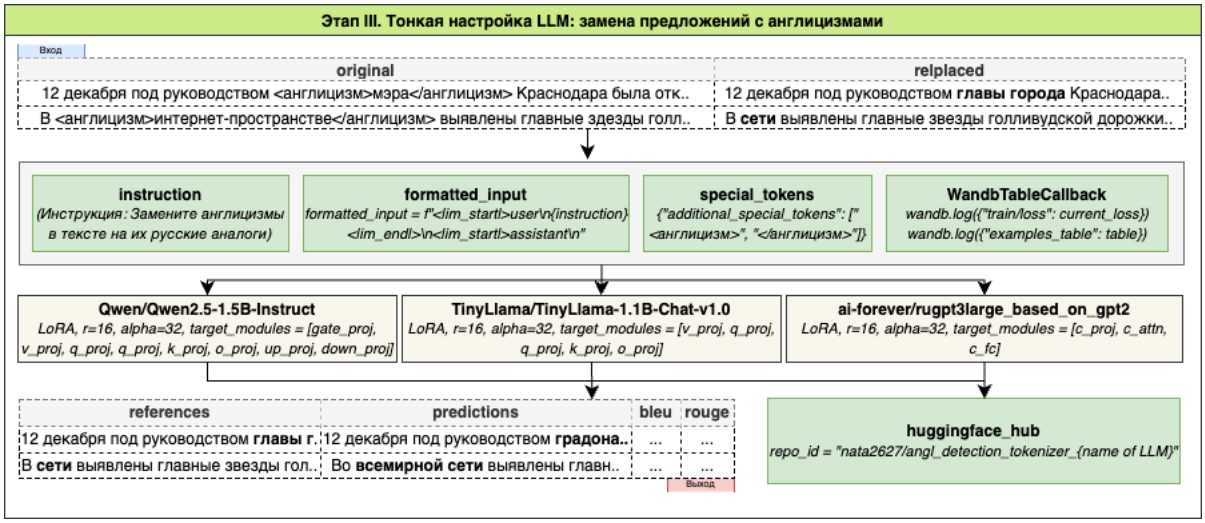


Рисунок 2.4.1. Архитектура экспериментальной части третьего этапа гибридной модели

Для эффективной адаптации предварительно обученных языковых моделей к задаче выявления англицизмов была применена комплексная методология тонкой настройки с использованием передовых техник оптимизации. Процесс обучения был реализован с применением

специализированных инструментов, обеспечивающих высокую эффективность и надежность результатов (таблица 2.4.1).

Таблица 2.4.1

Сравнение моделей Qwen/Qwen2.5-1.5B-Instruct, Qwen/Qwen2.5-3B-Instruct, Qwen/Qwen2.5-7B-Instruct, и YandexGPT-5-Lite-8B с учётом их числа параметров, требований к видеопамяти и производительности

Модель	Число параметров	Требования к видеопамяти (VRAM)	Особенности и производительность	Примечания
Qwen/Qwen2.5-1.5B-Instruct	1.5B	~12–16 ГБ VRAM	Хорошо подходит для ограниченных вычислительных мощностей, таких как Google Colab T4 (16 ГБ VRAM)	Подходит для Colab T4 и Kaggle
Qwen/Qwen2.5-3B-Instruct	3B	~16–24 ГБ VRAM	Требует больше памяти и вычислительных ресурсов, труднее запустить в средах с ограничениями (например, Colab T4)	Потребует распределённых вычислений при обучении
Qwen/Qwen2.5-7B-Instruct	7B	~24–40 ГБ VRAM	Требует серверных мощностей, возможно, несколько GPU с большой памятью	Используется на мощных серверных системах
YandexGPT-5-Lite-8B	8B	~40+ ГБ VRAM	Не загрузилась в Colab с GPU T4, требует гораздо больше VRAM для успешного выполнения	Очень тяжёлая модель для ограниченных ресурсов

Процесс тонкой настройки отобранных моделей реализовался с использованием инструментария библиотеки Hugging Face Transformers, предоставляющей унифицированный программный интерфейс для работы с различными архитектурами трансформеров и обеспечивающей стандартизированный подход к обучению и применению языковых моделей. Для оптимизации вычислительных затрат и повышения эффективности

обучения была имплементирована методология Parameter-Efficient Fine-Tuning (PEFT) с применением техники LoRA (Low-Rank Adaptation), обеспечивающей значительное сокращение количества обучаемых параметров модели через введение низкоранговых адаптеров в архитектуру.

Конфигурация LoRA для исследуемых моделей включала тщательно подобранные параметры, обеспечивающие оптимальное функционирование в рамках адаптации к специфической задаче замены англицизмов. Для модельного ряда Qwen2.5 (включая версии с 1.5B, 3B и 7B параметров) ранг матриц (r) был установлен на уровне 16, что обеспечивает эффективную компрессию параметрического пространства и снижение вычислительной сложности без существенного ущерба для точности результатов. Множитель масштабирования alpha со значением 32 усиливает влияние обучаемых параметров в процессе адаптации, обеспечивая более тонкую настройку модели на специфические особенности задачи замены англицизмов. Для предотвращения переобучения и повышения обобщающей способности модели использовался параметр dropout со значением 0.05, который случайным образом деактивирует часть нейронных связей в процессе обучения, способствуя формированию более робастной модели, устойчивой к вариативности входных данных.

Модель TinyLlama настраивалась с использованием аналогичных параметров ранга и масштабирования (16 и 32 соответственно), однако целевые слои для адаптации были ограничены компонентами o_proj, q_proj, v_proj и k_proj, что отражает структурные особенности данной архитектуры и специфику ее внутренней организации, отличающуюся от моделей семейства Qwen2.5.

Архитектура ai-forever/rugpt3large_based_on_gpt2 демонстрирует существенные структурные отличия от ранее рассмотренных моделей, что проявляется в специфическом наборе целевых слоев для адаптации: c_proj, c_attn и c_fc. Данная особенность обусловлена базированием модели на архитектуре GPT-2, характеризующейся иной внутренней организацией по

сравнению с более современными моделями семейства Qwen2.5 и TinyLlama, что требует специфической конфигурации параметров при тонкой настройке.

В таблице 2.4.2 отобразим конфигурации LoRA для различных моделей.

Таблица 2.4.2

Конфигурации LoRA для различных моделей

Модель	rank (r)	alpha	drop out	target_modules	base_model_name_or_path
Qwen2.5-1.5B-Instruct	16	32	0.05	["gate_proj", "v_proj", "q_proj", "k_proj", "o_proj", "up_proj", "down_proj"]	Qwen2.5-1.5B-Instruct
ai-forever/rugpt3large_based_on_gpt2	16	32	0.05	["c_proj", "c_attn", "c_fc"]	ai-forever/rugpt3large_based_on_gpt2
TinyLlama/TinyLlama-1.1B-Chat-v1.0	16	32	0.05	["o_proj", "q_proj", "v_proj", "k_proj"]	TinyLlama/TinyLlama-1.1B-Chat-v1.0

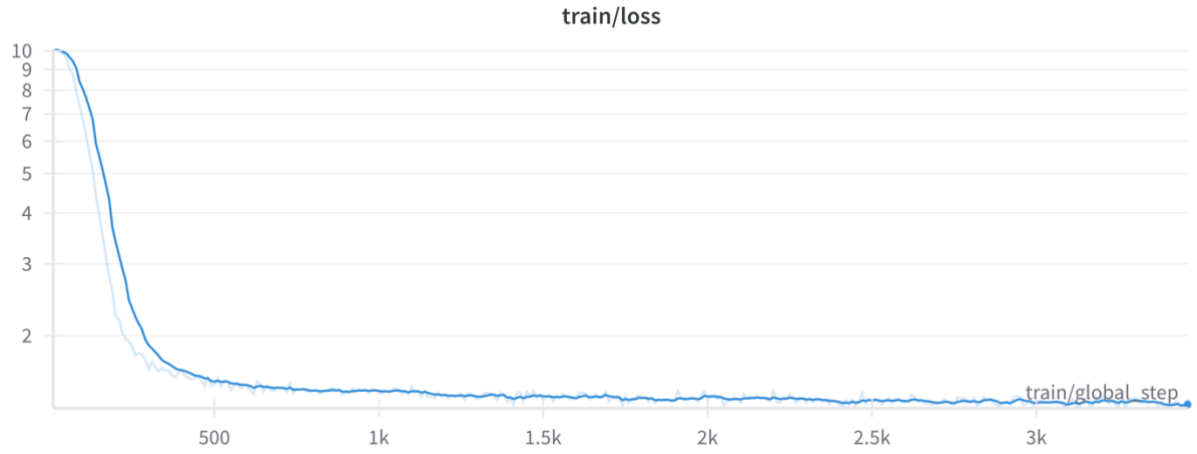
Для модели TinyLlama/TinyLlama-1.1B-Chat-v1.0 была дополнительно применена специализированная техника инициализации для расширения лексического словаря, что особенно важно при работе с задачами, требующими обработки специфической лексики, такой как англицизмы в русскоязычных текстах.

Для эффективного мониторинга и детального анализа процесса обучения была интегрирована специализированная платформа Weights & Biases (wandb), предоставляющая широкие возможности визуализации ключевых метрик в режиме реального времени, систематического отслеживания экспериментальных параметров и централизованного сохранения результатов.

Структура данных для обучения моделей представляла собой специализированный корпус текстов, содержащий пары предложений: оригинальные русскоязычные тексты с англицизмами и их эталонные варианты с заменой заимствований на семантически эквивалентные

русскоязычные аналоги. Общий объем тренировочного корпуса составил 18480 пар предложений, что обеспечивает репрезентативное покрытие различных контекстов использования англицизмов в русскоязычных текстах. Для валидационной оценки эффективности обучения было выделено 5% от общего объема данных, что составляет 116 примеров, позволяющих объективно оценить обобщающую способность модели и предотвратить переобучение на тренировочных данных.

Модель ai-forever/rugpt3large_based_on_gpt2 продемонстрировала монотонное снижение функции потерь с начального значения приблизительно 3.95 до финального показателя 1.39 на валидационной выборке. Характерной особенностью динамики обучения данной модели являются относительно высокие начальные значения потерь и постепенное, но стабильное их снижение на протяжении всего процесса обучения. К завершению обучения модель достигла значения функции потерь на тренировочной выборке 1.37, что свидетельствует о достаточной степени адаптации к поставленной задаче, однако эти показатели существенно уступают результатам других исследуемых архитектур (рисунок 2.4.2).



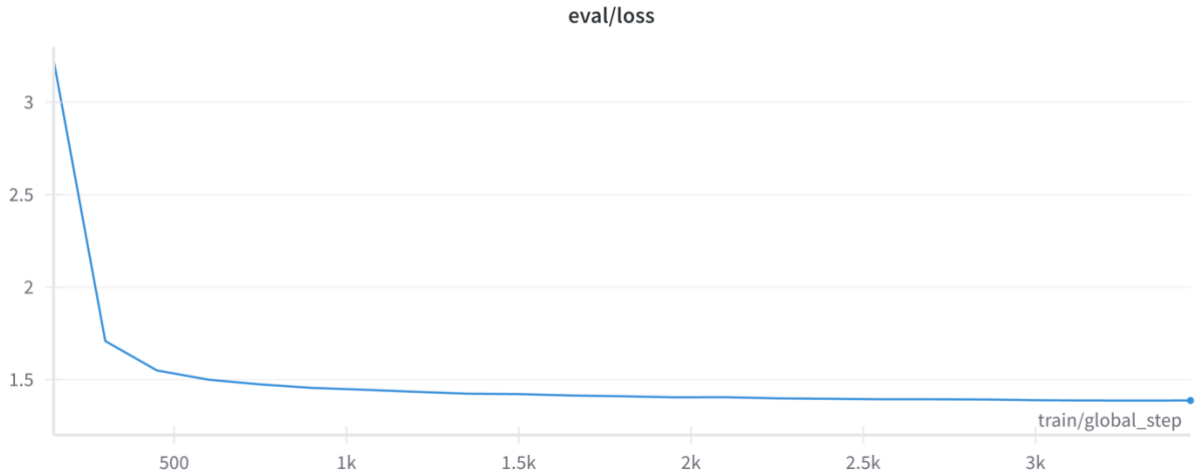


Рисунок 2.4.2. График функции потерь rugpt3large_based_on_gpt2 на а) тренировочной и б) валидационной выборках

Модель TinyLlama/TinyLlama-1.1B-Chat-v1.0, несмотря на меньшее количество параметров, продемонстрировала значительно более эффективную динамику обучения с существенно более низкими начальными значениями функции потерь (приблизительно 1.24 на валидационной выборке) и более быстрой сходимостью, достигнув значения 0.99 уже к середине процесса обучения. Финальное значение функции потерь составило 0.99 на валидационной выборке и 0.92 на тренировочной, что существенно превосходит показатели модели rugpt3large (рисунок 2.4.3).

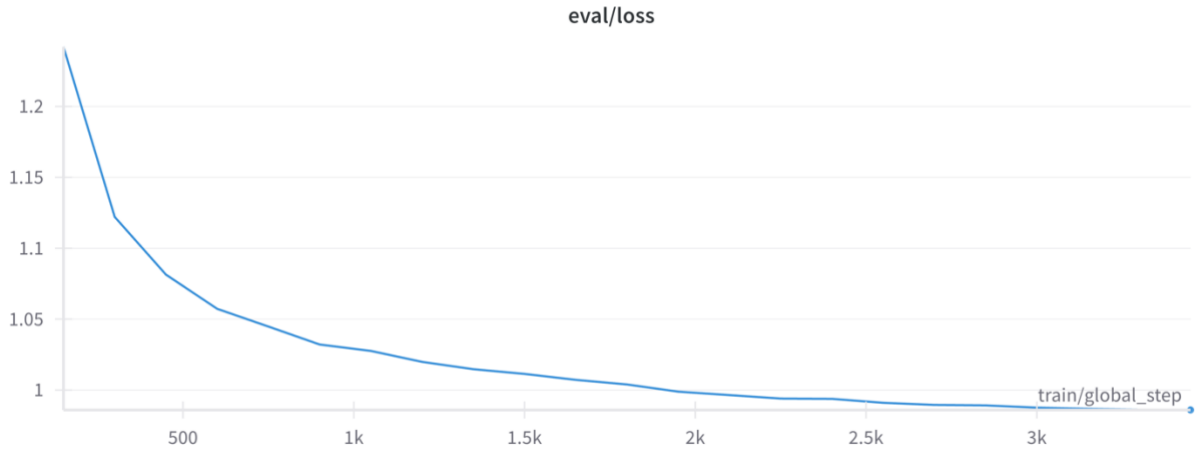
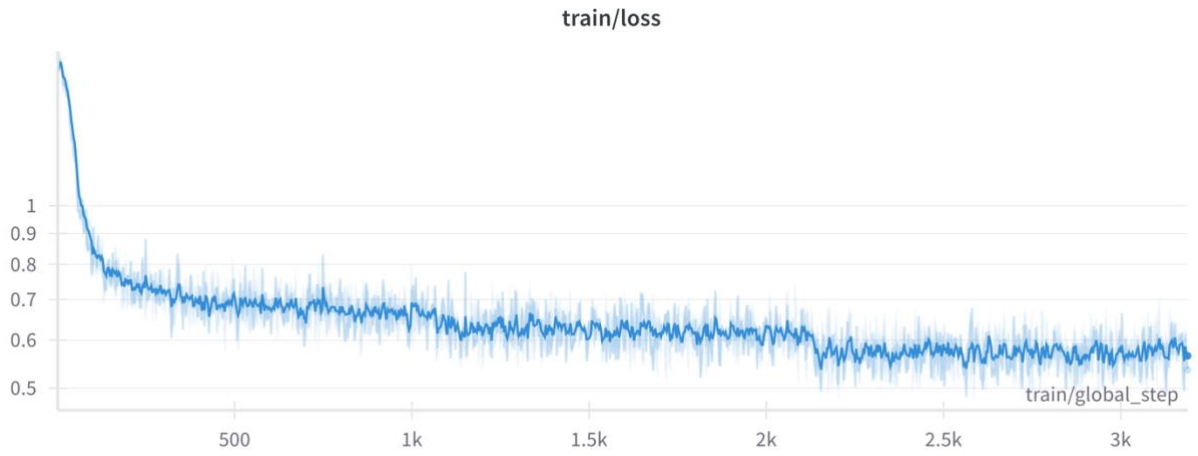


Рисунок 2.4.2. График функции потерь TinyLlama-1.1B-Chat-v1.0 на а) тренировочной и б) валидационной выборках

Модель Qwen/Qwen2.5-1.5B-Instruct продемонстрировала наиболее впечатляющие результаты с точки зрения минимизации функции потерь среди всех исследуемых архитектур. Начальное значение потерь на валидационной выборке составило всего 0.71, а финальный показатель достиг 0.63. На тренировочной выборке модель достигла значения функции потерь 0.54, что является наилучшим результатом среди всех тестируемых архитектур (рисунок 2.4.4).



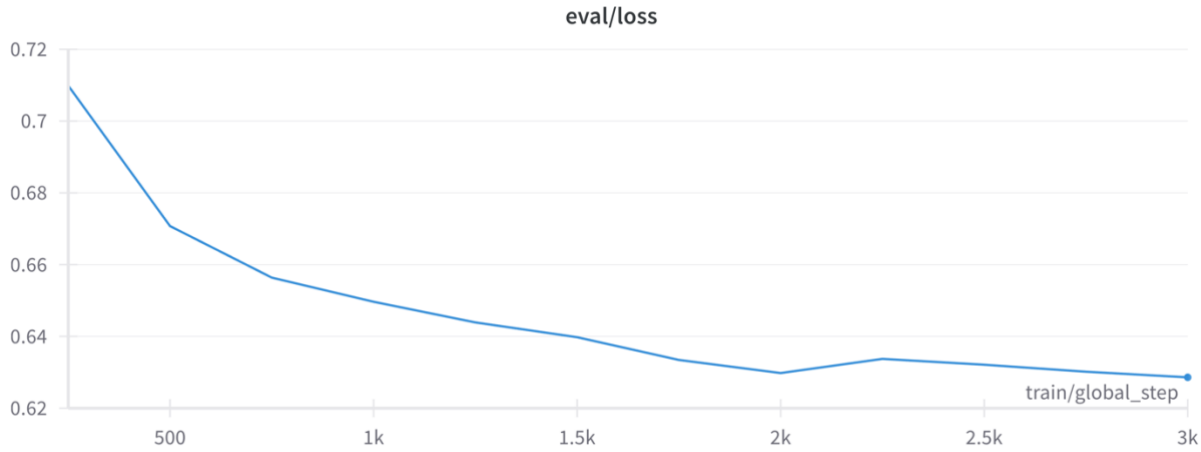


Рисунок 2.4.3. График функции потерь TinyLlama-1.1B-Chat-v1.0 на а) тренировочной и б) валидационной выборках

Анализ временных характеристик процесса обучения выявляет существенные различия в вычислительной эффективности исследуемых моделей. Общее время, затраченное на обучение, составило приблизительно 12307.72 секунды (около 3 часов 25 минут) для модели ai-forever/rugpt3large_based_on_gpt2, 12395.76 секунды (около 3 часов 26 минут) для модели TinyLlama/TinyLlama-1.1B-Chat-v1.0 и 23166.90 секунды (около 6 часов 26 минут) для модели Qwen/Qwen2.5-1.5B-Instruct (рисунок 2.4.3).

Таблица 2.4.3

Сравнительная таблица результатов обучения модели

Модель	Общее время обучения (сек)	Общее время обучения	Скорость обработки (примеров/сек)
ai-forever/rugpt3large_based_on_gpt2	12307,72	3 часа 25 минут	4,50
TinyLlama/TinyLlama-1.1B-Chat-v1.0	12395,76	3 часа 26 минут	4,47
Qwen/Qwen2.5-1.5B-Instruct	23166,90	6 часа 26 минут	2,20

Модель ai-forever/rugpt3large_based_on_gpt2 характеризуется наиболее стабильной динамикой снижения функции потерь без значительных флуктуаций, что свидетельствует о равномерном процессе обучения с

последовательным улучшением качества генерации. Визуализация градиентной нормы демонстрирует высокие значения на начальных этапах обучения с последующим быстрым снижением и стабилизацией, что соответствует типичному паттерну при использовании адаптивных методов оптимизации и свидетельствует о нормальном ходе процесса обучения.

Динамика качества генерации модели rugpt3large на примере задачи замены англицизма "рестрикции" на русскоязычный эквивалент в контексте санкционного давления демонстрирует постепенное, но неустойчивое улучшение результатов. На ранних этапах обучения (50-й шаг) наблюдаются существенные семантические отклонения от исходного содержания с полной потерей контекстуальной релевантности. Лишь к 3000-му шагу появляется семантически близкий эквивалент "запрет", а на 3450-м шаге наблюдается более адекватная замена "ограничения", хотя проблемы с орфографией и транслитерацией сохраняются.

В целом, процесс обучения этой модели показывает, что генеративные способности модели постепенно формируют связность и тематическую релевантность, однако остаются трудности с точностью терминологического замещения, сохранением именованных сущностей и соблюдением стилистической нормы (рисунок 2.4.4).

Таблица 2.4.4

Пример обучения rugpt3large_based_on_gpt2 на валидационном тексте

Шаг	Оригинал	Сгенерированное
50	Оригинал: Из организаций <англицизм>рестрикции</англицизм> коснулись оборонных компаний Elbit Systems, Parker Meggitt и Rafael UK. Таргет: Из организаций сдерживание коснулись оборонных компаний Elbit Systems, Parker Meggitt и Rafael UK	В этом случае не следует опасаться того, что ваши слова будут расценены как «антидемократические».
1000		Из организации взимания пошлины за оформление документов на вывоз товаров и услуг – отметили компании оборонные компании Elbit Consulting Services (ELC), ParkerMeggit and Rafel UK («Парк-Миггиттс») и RAFE (Research & Development).

2000		Из организации рестрикций коснулись оборонные компании Elbit Системс, Паркер Мегиттс и Rufael UK.
3000		Из организации запрета затронули оборонные компании Elbit Networks и Parker Meggitts (в России), а также Rafel International (США) и Rufael UK (Великобритания).
3450		Из организации ограничения коснулись оборонные компании Elbit Системы, Parkers Meg gt and Rafel UK.

Оценка качества генерации модели TinyLlama на примере научного текста о действии препарата "Фтортиазинон" демонстрирует интересную динамику адаптации к задаче замены англицизмов. На начальном этапе (50-й шаг) модель полностью теряет языковую релевантность, генерируя англоязычный текст вместо русскоязычного, что указывает на преобладание англоязычного распределения в начальных весах модели. Лишь к 3450-му шагу модель достигает полного соответствия с целевым текстом, включая корректную терминологию и замену англицизма на адекватный русскоязычный эквивалент (таблица 2.4.5).

Таблица 2.4.5

Пример обучения TinyLlama на валидационном тексте

Шаг	Оригинал/Таргет	Сгенерированное
50	Оригинал: Как объяснял академик, «Фтортиазинон» блокирует так называемую третью транспортную <англицизм>систему</англицизм>, которая помогает патогенным бактериям инфицировать клетки. Тагрет: Как объяснял академик, «Фтортиазинон» блокирует так называемую третью транспортную уклад, которая помогает патогенным	Text: The results showed that FTOrin, an anticancer agent, blocks the third transport system in the immune system, which allows pathogenic bacteria to infiltrate cancer cells.
1000		Текст: Как объяснял академик, «Фтортиазинон» блокирует так называемую третью транспортную путь, которая помогает патогенным бактериям инфицировать клетки.<|endoftext|>
1200		Как объяснял академик, «Фтортиазинон» блокирует так называемую третью транспортную пару, которая помогает патогенным бактериям инфицировать клетки.<|endoftext|></s>

3450	бактериям инфицировать клетки	Текст: Как объяснял академик, «Фтортиазинон» блокирует так называемую третью транспортную уклад, которая помогает патогенным бактериям инфицировать клетки.

Анализ качества генерации модели Qwen2.5-1.5B-Instruct на примере финансового текста с англицизмом "кредитов" демонстрирует наиболее эффективную динамику адаптации. Уже на 50-м шаге модель сохраняет общую структуру предложения, хотя замена англицизма еще не реализуется. К 1000-му шагу происходит переход к более релевантной лексике с использованием слова "заем" вместо "кредит", что свидетельствует о формировании адекватной стратегии лексического замещения. К 3150-му шагу наблюдается полное соответствие с целевым текстом, включая пунктуационную структуру и грамматическую норму, что свидетельствует о высокой эффективности модели в задаче замены англицизмов с сохранением семантической и стилистической целостности текста (рисунок 2.4.6)

Таблица 4.4.6

Пример обучения Qwen2.5 на валидационном тексте

Шаг	Оригинал	Сгенерированное
50	Оригинал: «Они в принципе запретили выдачи <англицизм>кредитов</англицизм> с высоким ПДН (показатель долговой нагрузки заемщика. Таргет: «Они в принципе запретили выдачу займов с высоким ПДН (показателем долговой нагрузки заемщика).»	«Они в принципе запретили выдачу кредитов с высоким ПДН (показатель долговой нагрузки заемщика.»
1000		Они в принципе запретили выдачу займа с высоким ПДН (показателем долговой нагрузки заемщика).
2000		«Они в принципе запретили выдачу займа с высоким ПДН (показателем долговой нагрузки заемщика).
3150		«Они в принципе запретили выдачу займа с высоким ПДН (показателем долговой нагрузки заемщика).»

В связи со значительным объемом обученных моделей (от 1.1 до 1.5 миллиардов параметров) было принято рациональное решение о сохранении

только весовых коэффициентов, полученных в результате тонкой настройки с использованием метода LoRA, вместо полных моделей.

Сравнительный анализ производительности моделей на основе метрик BLEU, ROUGE-1, ROUGE-2, ROUGE-L, а также средних длин предсказанных и эталонных текстов, демонстрирует значительные различия в качестве генерации (рисунок 2.4.4).

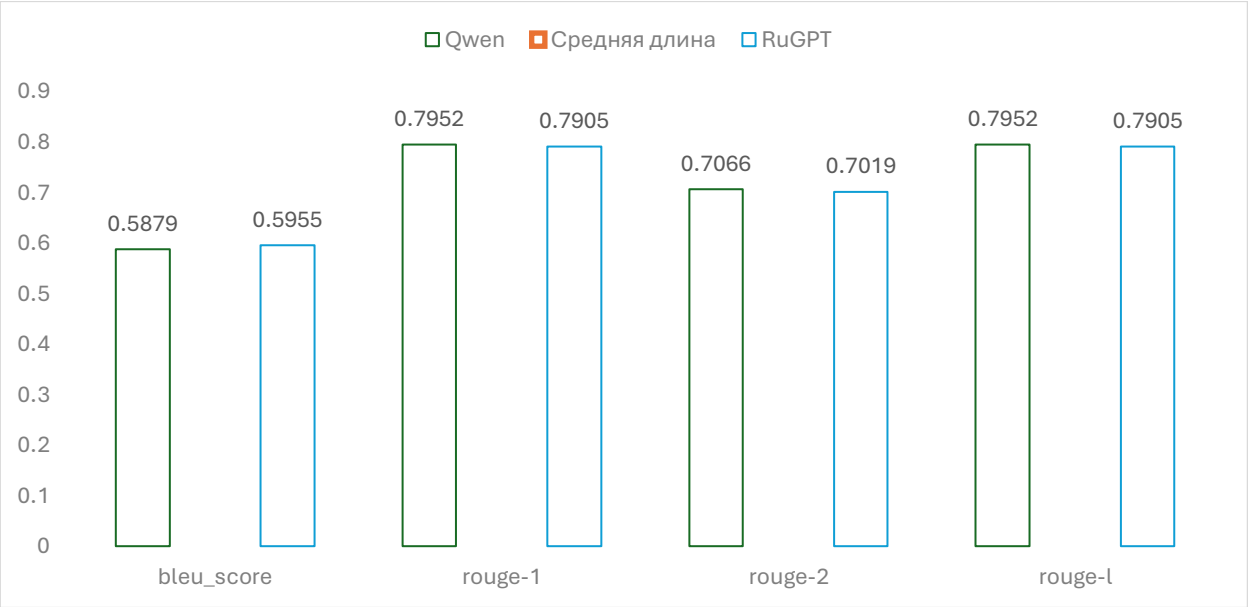


Рисунок 2.4.4. Метрики качества различных моделей на тестовых данных

Модель Llama показывает хорошие результаты по метрикам ROUGE-1, ROUGE-2 и ROUGE-L со значениями около 0.795, что свидетельствует о ее способности эффективно воспроизводить ключевые элементы эталонных текстов. Однако результаты по метрике BLEU существенно ниже (0.588), что может указывать на менее точное воспроизведение грамматической структуры и лексического состава эталонных текстов.

Модель Qwen демонстрирует наивысшие показатели по всем метрикам, особенно по BLEU (0.807) и ROUGE-1, ROUGE-2, ROUGE-L со значениями около 0.870. Это свидетельствует о высокой способности модели эффективно воспроизводить лексические и синтаксические особенности эталонных текстов, что делает ее оптимальным выбором для задач, требующих высокой

точности генерации. Средняя длина предсказанных текстов (17) несколько короче, чем у других моделей, что может указывать на тенденцию к генерации более лаконичных, но точных текстов, при этом средняя длина эталонных текстов (17.88) близка к данному значению, что свидетельствует о высокой адаптивности модели к различным стилям текста (рисунок 2.4.5).

Модель RuGPT показывает результаты, сходные с Llama по метрикам ROUGE-1, ROUGE-2 и ROUGE-L, но с несколько более низкими значениями, что указывает на ограниченную способность воспроизводить разнообразие и точность текста в сложных контекстах. Результаты по метрике BLEU (0.595) также свидетельствуют о менее точной генерации текста, что может быть связано с ограниченной адаптивностью модели к различным типам текстов. В отличие от Qwen, модель RuGPT генерирует более протяженные тексты (19.4 против 17 у Qwen), что может указывать на тенденцию к более подробным, но потенциально менее фокусированным результатам.

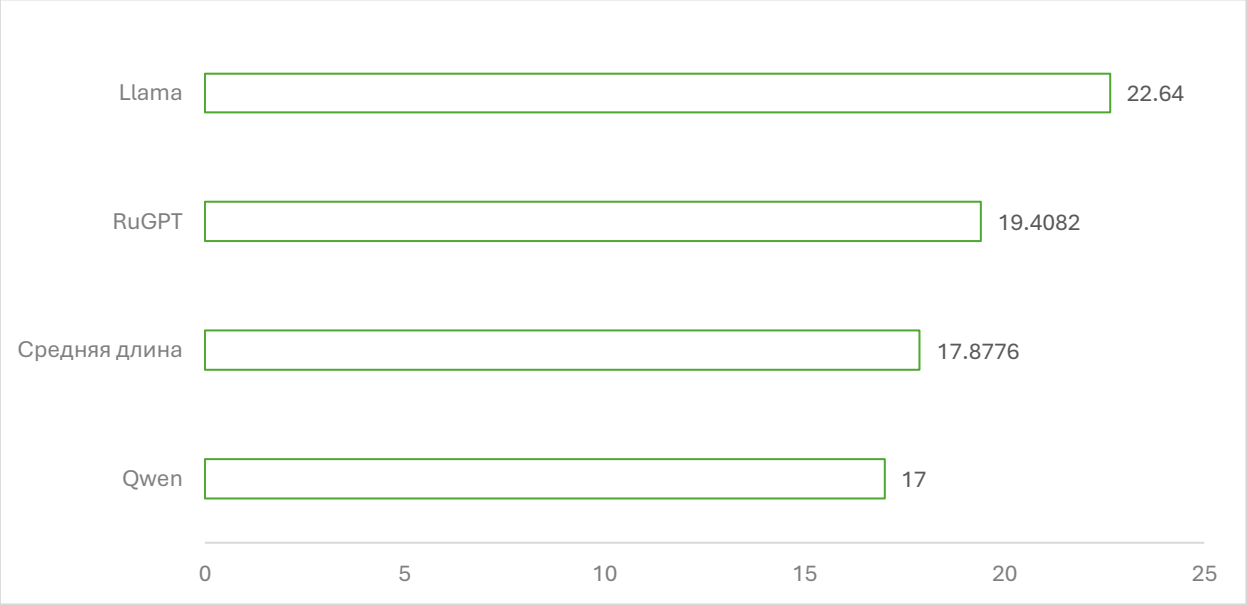


Рисунок 2.4.5. Сравнение среднего значения длины предложений на тестовых данных

Детальный анализ конкретных примеров генерации текста подтверждает наблюдаемые различия в эффективности моделей. Модель

Qwen демонстрирует практически идеальное соответствие предсказанных текстов эталонным образцам, достигая максимальных значений по метрикам exact_match и BLEU, а также близких к идеальным значений по метрикам ROUGE-1, ROUGE-2 и ROUGE-L. Модель Llama, несмотря на высокие показатели по метрикам ROUGE, имеет низкие значения exact_match, что указывает на наличие незначительных, но систематических различий между предсказанными и эталонными текстами. Модель RuGPT демонстрирует наименее эффективные результаты, особенно в случаях, требующих точной замены специфических терминов, таких как "инвестор" на "вкладчик средств".

Таким образом, сравнительный анализ различных моделей убедительно демонстрирует превосходство архитектуры Qwen/Qwen2.5-1.5B-Instruct для решения задачи автоматической замены англицизмов в русскоязычных текстах. Данная модель обеспечивает наивысшее качество генерации текста с максимальным сохранением семантической, грамматической и стилистической целостности исходного материала, что делает ее оптимальным выбором для практической реализации системы автоматической замены англицизмов, несмотря на более высокие вычислительные затраты по сравнению с альтернативными архитектурами.

Результаты экспериментального исследования убедительно доказывают эффективность разработанной гибридной модели для автоматического выявления и замены англицизмов в русскоязычных новостных текстах. Трехэтапный подход позволил преодолеть ограничения существующих решений, обеспечив высокую точность на всех стадиях обработки текста. Особую значимость представляет выявленная оптимальная конфигурация технологического стека: XGBoost для классификации, Qwen-3B для генерации синонимов и Qwen2.5-1.5B-Instruct для финальной обработки. Принципиально важным является обнаруженный баланс между параметрической сложностью моделей и качеством обработки текста — увеличение числа параметров свыше определенного порога не приводит к существенному улучшению результатов.

Раздел 3. Тестирование гибридной модели и обсуждение результатов

3.1 Построение гибридной модели на основе результатов проведенных экспериментов

После проведения сравнительного анализа различных методов машинного обучения для задачи классификации англицизмов, была выбрана модель градиентного бустинга XGBoost как оптимальное решение. Данный выбор обусловлен выдающимися результатами модели по всем ключевым метрикам эффективности, продемонстрированными в ходе сравнительного анализа различных алгоритмов машинного обучения. Архитектура XGBoost (eXtreme Gradient Boosting) представляет собой усовершенствованную реализацию градиентного бустинга с рядом оптимизаций, обеспечивающих повышенную производительность и точность классификации.

Ключевые преимущества выбранной модели XGBoost включают: эффективные механизмы L1 и L2 регуляризации, предотвращающие переобучение и повышающие обобщающую способность модели; оптимизированные алгоритмы для работы с разреженными матрицами признаков, характерными для задач обработки естественного языка; возможность распараллеливания вычислительных процессов при построении деревьев решений, что повышает вычислительную эффективность модели; встроенные механизмы обработки пропущенных значений в данных, обеспечивающие устойчивость к неполноте информации; функциональность ранней остановки, позволяющую прекращать обучение при отсутствии улучшений на валидационной выборке, что предотвращает переобучение и оптимизирует вычислительные затраты.

Модель XGBoost была обучена на масштабном корпусе данных, включающем 100,000 примеров, что обеспечивает репрезентативное покрытие различных контекстов использования англицизмов в русскоязычных текстах. Процесс обучения проводился на 90,000 примерах, а тестирование эффективности модели — на 10,000 примерах, что создает основу для

статистически значимой оценки качества классификации и обобщающей способности модели.

Анализ матрицы ошибок демонстрирует исключительную точность классификации: из 10,000 тестовых примеров 4,969 неанглицизмов были корректно классифицированы как неанглицизмы (истинно отрицательные результаты, TN), 4,980 англицизмов были правильно определены как англицизмы (истинно положительные результаты, TP), 31 неанглицизм был ошибочно классифицирован как англицизм (ложноположительные результаты, FP), и только 20 англицизмов не были распознаны моделью (ложноотрицательные результаты, FN). Общее количество ошибок составляет всего 51 из 10,000 примеров, что подтверждает исключительную точность модели и ее высокую практическую применимость для задачи идентификации англицизмов в русскоязычных текстах (рисунок 3.1.1).

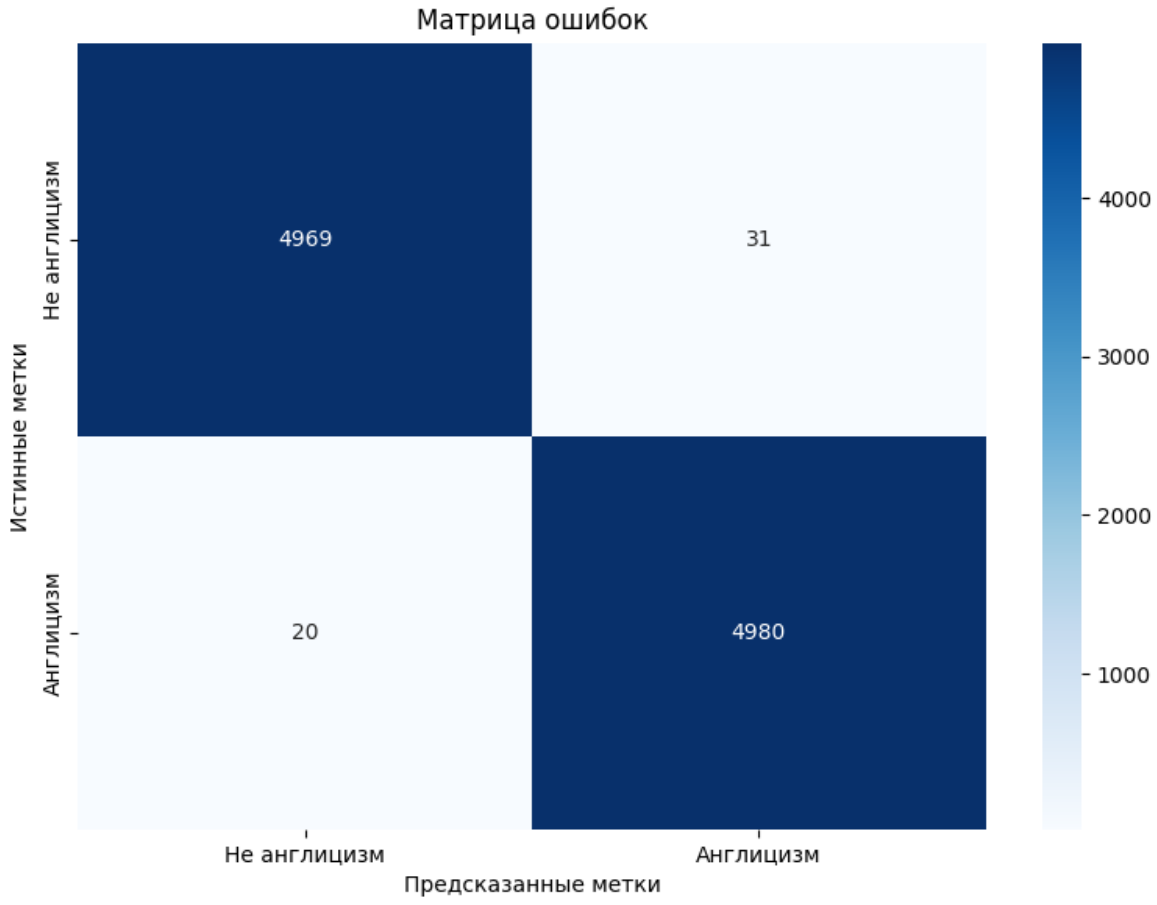


Рисунок 3.1.1. Матрица ошибок

ROC-кривая модели демонстрирует практически идеальную форму с площадью под кривой (AUC) 0.999, что является наивысшим возможным показателем для классификационной модели и свидетельствует о превосходной способности алгоритма различать англицизмы и неанглицизмы при различных пороговых значениях вероятности (рисунок 3.1.2). Такой высокий показатель AUC указывает на оптимальное соотношение между чувствительностью и специфичностью модели, что критически важно для практического применения в задачах классификации лексических единиц.

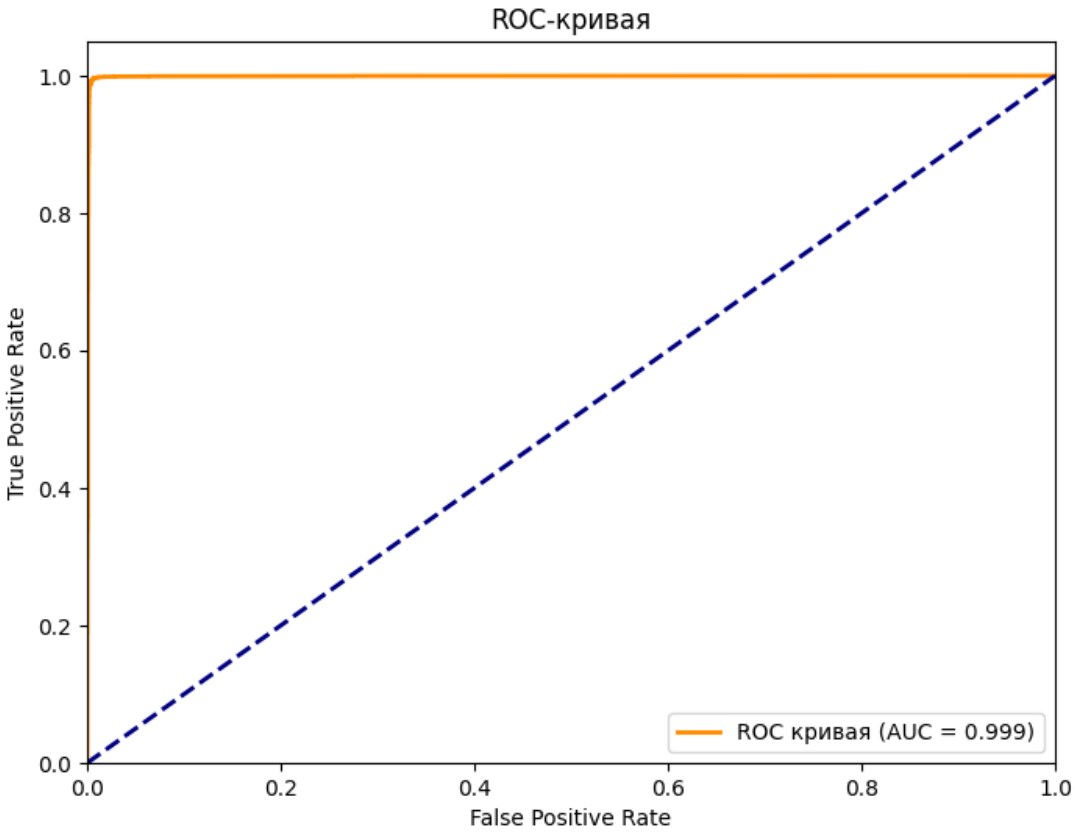


Рисунок 3.1.2. ROC-кривая

Анализ динамики обучения модели XGBoost позволяет выявить ряд ключевых особенностей процесса формирования классификатора: быстрое начальное снижение значений функции потерь и ошибок классификации в

первые 20-40 итераций, что свидетельствует о высокой скорости сходимости алгоритма на начальных этапах обучения; постепенное замедление улучшений после 40-60 итераций, что соответствует типичному паттерну обучения моделей машинного обучения, когда дальнейшие улучшения требуют более тонкой настройки параметров; очень близкие значения функции потерь на тренировочной и тестовой выборках, что указывает на отсутствие переобучения и высокую обобщающую способность модели; небольшое расхождение между тренировочной и тестовой ошибками после 60 итераций, что является нормальным явлением и не свидетельствует о значительном переобучении модели. Финальное значение ошибки классификации на тестовой выборке составляет приблизительно 0.005 (0.5%), что соответствует точности 0.995 и подтверждает исключительно высокую эффективность модели в решении задачи идентификации англицизмов.

Сравнительный анализ результатов финальной модели XGBoost, обученной на полном корпусе данных (100,000 примеров), с результатами предварительных экспериментов на меньшей выборке (20,000 примеров) демонстрирует значительное улучшение всех метрик эффективности. Наиболее существенное улучшение наблюдается в показателе полноты (Recall), что свидетельствует о том, что расширенный корпус данных позволил модели более эффективно выявлять англицизмы различных типов, включая редкие или нетипичные случаи их употребления. Данное наблюдение подтверждает важность формирования репрезентативного и масштабного корпуса обучающих данных для достижения высокой эффективности моделей машинного обучения в задачах обработки естественного языка (таблица 3.1.1):

Таблица 3.1.1

Сравнение с предыдущими результатами

Метрика	XGBoost (20k примеров)	XGBoost (100k примеров)	Улучшение
Accuracy	0.9885	0.9949	+0.0064
Precision	0.9929	0.9938	+0.0009
Recall	0.9840	0.9960	+0.0120
F1-мера	0.9884	0.9949	+0.0065

На втором этапе разработки комплексного решения для автоматической замены англицизмов в качестве оптимальной языковой модели была выбрана архитектура QWEN2.5-3B-Instruct, обеспечивающая оптимальный баланс между скоростью обработки текстовой информации и качеством генерации русскоязычных эквивалентов англицизмов. Данный выбор основан на результатах комплексного сравнительного анализа различных языковых моделей, представленного в предыдущих разделах исследования, и учитывает как качественные показатели эффективности, так и вычислительные требования различных архитектур. Рассмотрим пример работы модели:

1. На вход модель получила предложение «Поводов для эвакуации жителей Харькова пока нет, заявил Игорь Терехов в интервью украинскому изданию LIGA», а также список англицизмов в этом предложении [«интервью»].

2. Модель предложила следующие варианты для замены слова: ["разговор", "беседа", "переговор", "допрос", "отчет", "дискуссия", "встреч", "монолог"].

3. Среди них были выбраны 3 наиболее подходящие замены (по семантическому сходству: [«разговор», «беседа», «дискуссия»].

4. Были составлены 3 предложения с этими словами:

А) Поводов для эвакуации жителей Харькова пока нет, заявил Игорь Терехов в беседе с украинским изданием LIGA»

Б) Поводов для эвакуации жителей Харькова пока нет, заявил Игорь Терехов в разговоре с украинским изданием LIGA»

В) Поводов для эвакуации жителей Харькова пока нет, заявил Игорь Терехов во время дискуссии с украинским изданием LIGA»

5. Среди полученных предложений было выбрано одно – с наибольшей семантической схожестью, им оказался вариант А.

Визуальное представление полного процесса генерации синонимов с учетом семантического сходства представлено на рисунке 2.3.1:

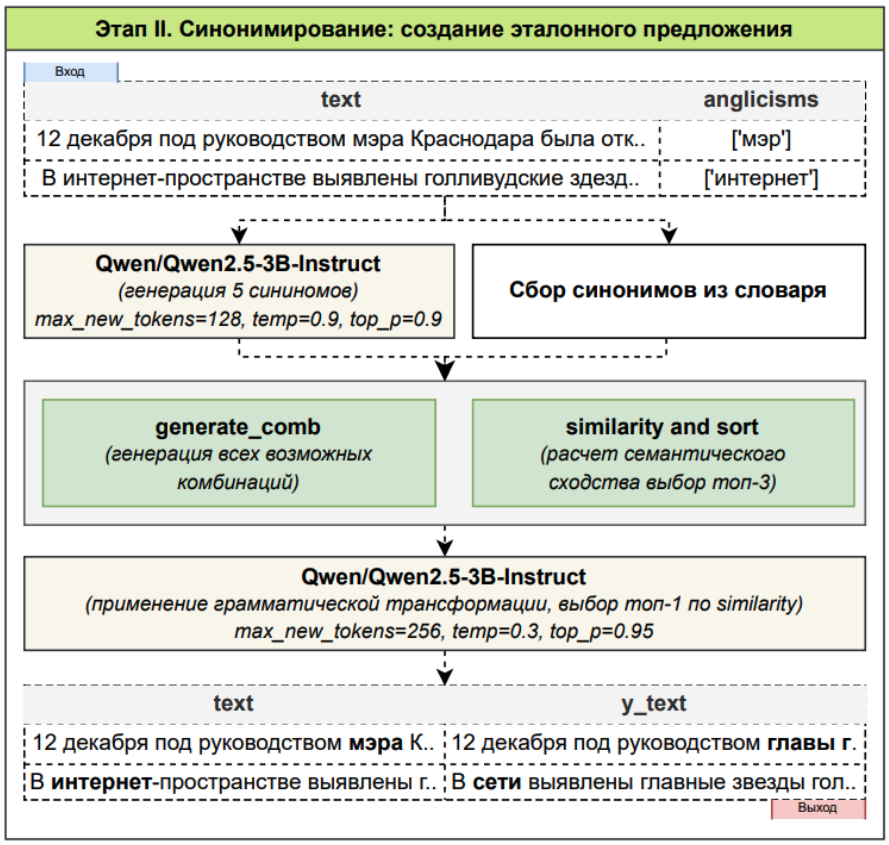


Рисунок 3.1.3. Процесс генерации и отбора синонимов для англицизмов

Данная архитектура наглядно демонстрирует реализацию процесса замены англицизмов с сохранением семантической и грамматической целостности текста, соответствуя описанной в параграфе методологии, включающей генерацию синонимов, семантическую оценку и грамматическую нормализацию.

Таким образом, использование многоступенчатого подхода к генерации текста без англицизмов позволяет, во-первых, гарантировать получения достаточного количества наиболее подходящих синонимов, во-вторых, исключить попадание других англицизмов в текст, в-третьих, эффективно использовать вычислительные ресурсы, путем избегания необходимости генерации всех доступных вариантов замен, в-четвертых, значительно повысить качество итогового предложения, путем сравнения трех лучших вариантов.

Для реализации третьего этапа было принято решение выбрать модель Qwen2.5-1.5B-Instruct, учитывая как теоретические, так и практические аспекты, выявленные в ходе проведенных экспериментов.

Трансформерные архитектуры, к которым относится Qwen2.5-1.5B-Instruct, демонстрируют выдающиеся показатели в области генерации текста, перевода и решения комплексных задач, требующих контекстуального понимания и мультифункциональности. Обоснованность данного выбора определяется рядом ключевых характеристик, среди которых особенно выделяется способность модели адаптироваться к специфическим инструкциям (Instruct), предоставляя значительное преимущество при выполнении задач, ориентированных на точность и высококачественную текстовую генерацию. Практический опыт применения Qwen2.5-1.5B-Instruct свидетельствует о ее высокой адаптивности к разнообразным запросам, делая ее универсальным инструментом для широкого спектра приложений – от создания контента до специализированных аналитических и вопросно-ответных систем.

Начальный этап работы с Qwen2.5-1.5B-Instruct включает загрузку предварительно обученной модели совместно с соответствующим токенизатором. Данная процедура реализуется посредством метода AutoPeftModelForCausalLM.from_pretrained с указанием таких параметров, как device_map="auto", обеспечивающего автоматический выбор оптимального вычислительного устройства, и torch_dtype=torch.float16, позволяющего ускорить вычисления за счет использования полуточных чисел. Загрузка токенизатора осуществляется через метод AutoTokenizer.from_pretrained, необходимый для преобразования текстовой информации в понятный для модели формат токенов (рисунок 3.1.4).

```python
# Функция загрузки модели
def load_model2():
    model = AutoPeftModelForCausalLM.from_pretrained(
        "nata2627/angl_detection_tokenizer_qwen",
        device_map="auto",
        torch_dtype=torch.float16,
        ignore_mismatched_sizes=True
    )
    tokenizer = AutoTokenizer.from_pretrained("nata2627/angl_detection_tokenizer_qwen")

    model = model.to("cuda" if torch.cuda.is_available() else "cpu")
    model.eval()
    return model, tokenizer
```

Рисунок 3.1.4. Код загрузки дообученной автором модели и соответствующего токенизатора

После успешной загрузки компонентов модель перемещается на доступное вычислительное устройство и переводится в режим оценки посредством метода .eval(). Указанная операция имеет принципиальное значение, поскольку в режиме оценки функционирование модели отличается от обучающего режима благодаря отключению таких механизмов, как dropout, что обеспечивает стабильность работы при тестировании системы.

Следующий этап алгоритма предполагает получение моделью входного предложения для последующей обработки. В качестве иллюстративного примера выступает предложение, содержащее англицизмы, требующие замены на русскоязычные эквиваленты. Данное предложение передается в модель для дальнейшей обработки согласно заданным параметрам.

Процесс генерации начинается с формирования специализированной инструкции, определяющей специфику обработки текста моделью. Инструкция содержит четкие указания относительно необходимых трансформаций исходного текста, в данном контексте – замены англицизмов на русские аналоги. Текст вместе с инструкцией форматируется в соответствии с требованиями модели, включая использование специальных разграничительных тегов <|im_start|>user и <|im_end|>, обеспечивающих четкую демаркацию пользовательского ввода и ассистентского ответа (рисунок 3.1.5).

```python
# Генерация предсказаний с правильным форматированием промпта
predictions = []
for input_text in original:
    # Создаем инструкцию для модели (как во время обучения)
    instruction = (
        "Инструкция: Замените англицизмы в тексте на их русские аналоги.\n\n"
        f"Текст: {input_text}\n\n"
        f"Результат:"
    )

    # Форматируем промпт как во время обучения
    formatted_input = f"<|im_start|>user\n{instruction}<|im_end|>\n<|im_start|>assistant\n"

    # Токенизация
    inputs = tokenizer2(formatted_input, return_tensors="pt").to(model2.device)

    # Генерация с теми же параметрами, что и в обучении
    with torch.no_grad():
        outputs = model2.generate(
            **inputs,
            max_new_tokens=256,
            temperature=0.7,
            top_p=0.9,
            do_sample=True,
            pad_token_id=tokenizer2.eos_token_id
        )

    # Декодирование и извлечение только ответа ассистента
    pred_text = tokenizer2.decode(outputs[0], skip_special_tokens=False)
    assistant_part = pred_text.split("<|im_start|>assistant\n")[-1].split("<|im_end|>")[0]
    predictions.append(assistant_part.strip())
```

Рисунок 3.1.5. Код загрузки генерации предсказаний с форматированием промта

Сформированный текст подвергается токенизации, после чего полученные токены направляются в модель. Генерация ответа реализуется через метод model.generate с настройкой различных параметров, включая max_new_tokens для лимитирования длины ответа, temperature для регулирования стохастичности генерации, top_p для управления вероятностным распределением при выборке, а также других параметров, влияющих на качественные характеристики и вариативность генерируемого контента.

По завершении генерации ответа производится его декодирование в текстовый формат посредством метода tokenizer.decode. Критически важным элементом данного этапа выступает извлечение релевантной части текста, относящейся непосредственно к ответу ассистента, с исключением служебных

тегов и метаданных. Достижение этой цели обеспечивается через сегментацию строки для выделения целевого фрагмента информации.

Заключительная фаза алгоритма предусматривает вычисление метрик качества моделирования, таких как BLEU и ROUGE, посредством функционала библиотеки datasets. Указанные метрики позволяют количественно оценить степень соответствия сгенерированного текста эталонным образцам, предоставляя объективную характеристику качества генерации на основе вычисления показателей семантической и структурной близости между сравниваемыми текстами.

Представленная последовательность этапов формирует комплексный протокол взаимодействия с моделью, охватывающий полный цикл от инициализации и конфигурирования до генерации текстового контента и его качественной оценки посредством специализированных метрических инструментов.

3.2 Оценка качества работы модели на тестовом наборе данных

Для оценки эффективности предложенной трехэтапной гибридной модели детекции и замены англицизмов был реализован инференс на разнородных тестовых материалах. Экспериментальная выборка охватывала как новостные тексты, соответствующие характеристикам обучающего корпуса, так и разговорные высказывания, представляющие принципиально новый тип данных для системы. Данный методологический подход позволил провести всестороннюю оценку генерализационных возможностей и адаптивности разработанного алгоритмического комплекса.

Архитектурная концепция гибридной модели предусматривает реализацию трех последовательных функциональных этапов, схематически представленных на рисунке 3.2.1. Каждый структурный компонент выполняет специфические операции в контексте лингвистической обработки текстов, содержащих англоязычные заимствования.

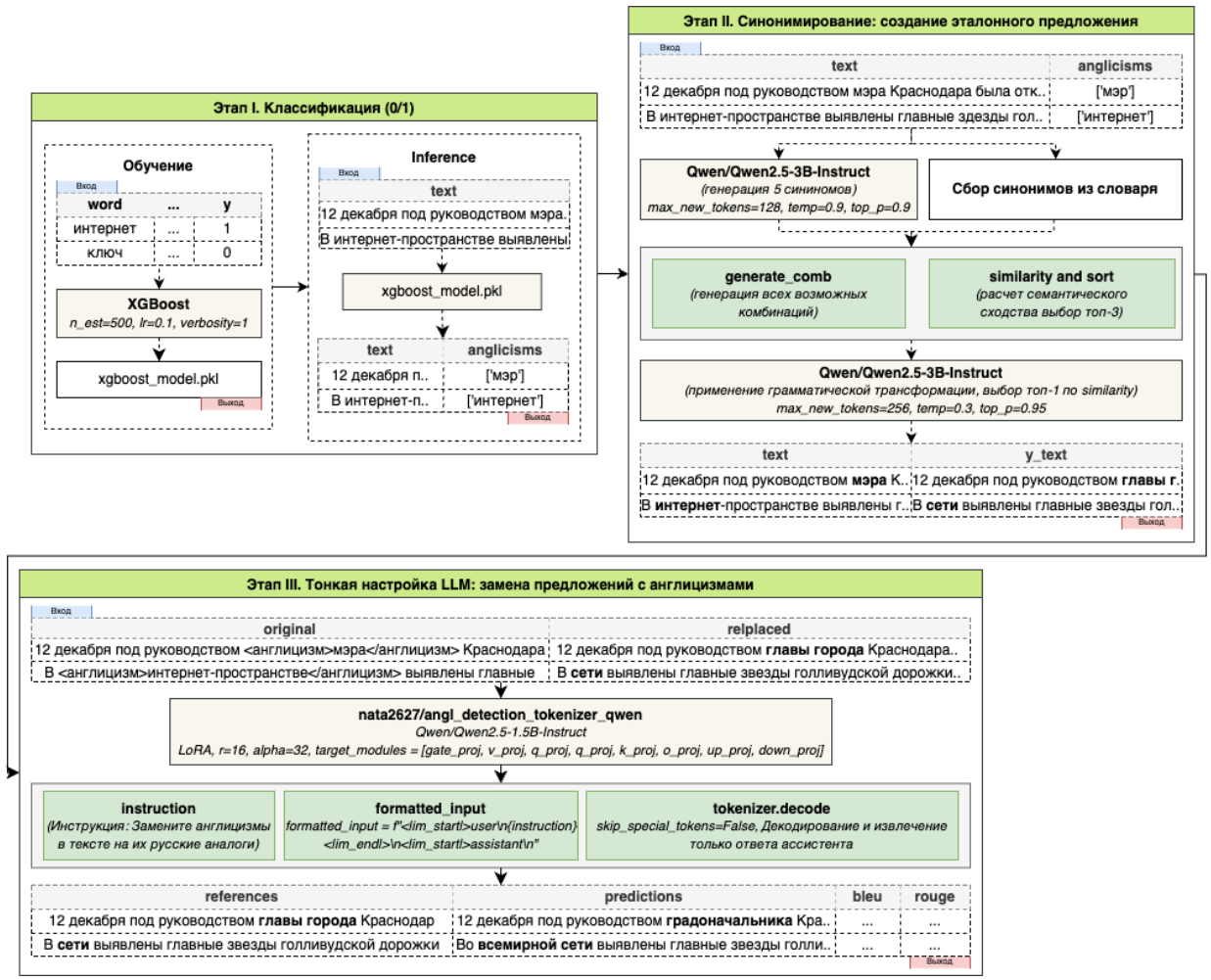


Рисунок 3.2.1 Архитектура модели и этапы обработки

Начальный этап системы – классификация лексических единиц – реализован в верхней левой секции схемы. Процессуальный цикл инициируется обучением алгоритмов машинного обучения на аннотированном корпусе текстов с датировкой 12 декабря. Архитектурная структура интегрирует предварительную обработку текстовой информации через токенизацию, экстракцию многомерных признаков (морфологических, контекстуальных, статистических, синтаксических) и обучение трех дифференцированных моделей: линейной регрессии, случайного леса и XGBoost. Параллельно функционирует лексикографический механизм с обращением к специализированному словарю англицизмов. Результирующим выходом этапа становится бинарная категоризация лексем как англоязычных заимствований либо исконно русских слов.

Вторая фаза алгоритма – синонимизация с формированием эталонных фраз – локализована в правой верхней части диаграммы. Входными параметрами здесь выступают текстовые фрагменты с идентифицированными англицизмами. Генерация русскоязычных синонимических эквивалентов осуществляется посредством нейросетевой архитектуры Qwen/Qwen2.5-3B-Instruct с применением следующих параметрических настроек: max_new_tokens=128, temp=0.9, top_p=0.9. Дополнительным ресурсом выступает предварительно сформированный синонимический словарь. Последующие алгоритмические блоки включают generate_comb, обеспечивающий формирование полного спектра комбинаторных замен, и similarity and sort, реализующий квантитативную оценку семантической когерентности с ранжированием вариантов по степени приближенности к исходному тексту. Финальная лингвистическая нормализация производится идентичной моделью с редуцированным параметром temperature=0.3 для обеспечения грамматической корректности генерируемых конструкций. Итоговым продуктом этапа становится референсное предложение с заменой англицизмов на русскоязычные лексические аналоги.

Заключительный этап алгоритмического комплекса – точная настройка языковой модели – визуализирован в нижней части схематического представления. На данной стадии используются информационные массивы, сформированные на предыдущих этапах: оригинальные текстовые конструкции с англицизмами и референсные предложения с их русскоязычными субститутами. Процессуальный цикл включает компоновку тренировочного корпуса в инструктивном формате: instruction (директива "Замени англицизмы в тексте на их русские аналоги"), tokenized_input (токенизированный входной массив с англицизмами), output_tokens (целевой выход с заменой заимствований) и translatedInEnglisch (верификационный механизм через англоязычную транспозицию). Обучение нейросетевой архитектуры реализуется на основе моделей Qwen-Instruct-3.5B и TinyLlama-1.1B-Chat с интеграцией через программный интерфейс huggingface_hub.

Эвалюационная методология модели на тестовых данных базировалась на комплексе метрических показателей, включая BLEU (Bilingual Evaluation Understudy), квантифицирующий корреляционные связи между модельным выходом и эталонным текстом, коэффициент ROUGE (Recall-Oriented Understudy for Gisting Evaluation), лексическую диверсификацию выходных конструкций, квантитативные характеристики успешно замененных англицизмов и сохранение семантической целостности исходного высказывания, оцениваемое лингвистическими экспертами.

Сравнительный анализ проводился между финальными генерациями модели и референсными предложениями, созданными профессиональными лингвистами. Особое внимание исследователей фокусировалось на процессуальной обработке разговорных конструкций, представляющих собой новый лингвистический паттерн для системы.

Новостные тексты, будучи типологически близкими к тренировочному корпусу, априори предполагали более высокую эффективность алгоритмической обработки. В таблице 3.2.1 представлена поэтапная трансформация новостных высказываний в процессе модельной обработки.

Таблица 3.2.1

Примеры трансформации новостных предложений на каждом этапе модели

Исходное предложение	Этап 1 (Детекция англицизмов)	Этап 2 (Создание эталонного предложения)	Этап 3 (Финальное преобразование)
"Новый стартап получил крупные инвестиции от известного бизнесмена, что позволит компании выйти на международный рынок."	"Новый стартап получил крупные инвестиции от известного бизнесмена, что позволит компании выйти на международный рынок.","[""стартап"", ""бизнесмена"",""инвестиции"", ""компании"", ""международный""]"	"Новое молодое предприятие получило крупные вложения от известного предпринимателя, что позволит организации выйти на мировой рынок."	"Новое предприятие получило крупную вкладывательность средств от известного предпринимателя, что позволит такое объединение выйти на

			межнациональный рынок."
"На пресс-конференции топ-менеджер презентовал дорожную карту развития компании, подчеркнув важность инноваций."	"На пресс-конференции топ-менеджер презентовал дорожную карту развития компании, подчеркнув важность инноваций."	"На пресс-конференции топ-менеджер презентовал дорожную карту развития компании, подчеркнув важность инноваций."	"На пресс-конференции топ-менеджер презентовал дорожную карту развития компании, подчеркнув важность инноваций."
"Четырехкратный чемпион «Формулы-1» Макс Ферстаппен сообщил в своем Instagram (принадлежит Meta, компания признана экстремистской и запрещена в России), что впервые стал отцом."	"Четырехкратный чемпион «Формулы-1» Макс Ферстаппен сообщил в своем Instagram (принадлежит Meta, компания признана экстремистской и запрещена в России), что впервые стал отцом.","[""чемпион""]	"Четырехкратный победитель «Формулы-1» Макс Ферстаппен сообщил в своем Инстаграме (принадлежащем Мете, организации, признанной экстремистской и запрещенной в России), что впервые стал отцом."	"Четырехкратный победитель «Формулы-1» Макс Ферстаппен сообщил в своем Instagram (принадлежащем Meta, компания признанной экстремистской и запрещенной в России), что впервые стал отцом."

Аналитическая интерпретация результатов свидетельствует о высокой эффективности модели в детекции англицизмов в новостном дискурсе. На первом этапе успешно идентифицируются как эксплицитные заимствования (стартап, бизнесмен), так и лексические единицы, давно ассимилированные русским языком (компания, международный), что демонстрирует высокую сенситивность алгоритма XGBoost. Второй этап характеризуется способностью модели Qwen2.5-3B-Instruct генерировать семантически когерентные конструкции с заменой англицизмов на русскоязычные эквиваленты. Третья фаза в отдельных случаях демонстрирует нетривиальные результаты, дивергирующие от эталонных предложений, но сохраняющие семантическую целостность высказывания.

Особый исследовательский интерес представляет анализ алгоритмической эффективности при обработке разговорных высказываний,

существенно отличающихся от новостного дискурса по структурно-синтаксическим, лексическим и стилистическим параметрам. Таблица 3.2.2 иллюстрирует поэтапную трансформацию разговорных конструкций в процессе модельной обработки.

Таблица 3.2.2

Примеры трансформации разговорных предложений на каждом этапе модели

Исходное предложение	Этап 1 (Детекция англицизмов)	Этап 2 (Создание эталонного предложения)	Этап 3 (Финальное преобразование)
"Скинь мне лук твоего аутфита для вечеринки."	"Скинь мне лук твоего аутфита для вечеринки.","[""скинь"", ""лук"", ""аутфита""]"	"Отправь мне изображение твоего наряда для вечеринки."	"Отправь мне снимок твоей одежды для вечеринки."
"Блогер опубликовал пост, вызвавший бурный фидбек в социальных сетях."	"Блогер опубликовал пост, вызвавший бурный фидбек в социальных сетях.","[""блогер"",""пост"", ""социальных""]"	"Создатель сетевого дневника разместил заметку, вызвавшую бурный отклик в сетях общения."	"Автор публикации разместил запись, вызвавшую бурный отклик в общественной сети."
"Извиняюсь, я опоздал, был на митинге с заказчиком."	"Извиняюсь, я опоздал, был на митинге с заказчиком.","[""митинге""]"	"Прошу прощения, я опоздал, был на собрании с заказчиком."	"Извиняюсь, я опоздал, был на совещании с заказчиком."

При интерпретации результатов обработки разговорных высказываний выделяются несколько ключевых особенностей. Во-первых, система демонстрирует высокую детектирующую способность в отношении специфичных для разговорной коммуникации англицизмов, таких как "скинь", "лук", "аутфит", "фидбек". Данные лексические единицы широко распространены в молодежном социолекте и цифровой коммуникации, но редко встречаются в официальных информационных источниках.

Во-вторых, на втором процессуальном этапе нейросетевая архитектура Qwen2.5-3B-Instruct формирует естественные лексические субституты англицизмов с сохранением семантической целостности высказывания.

Например, трансформация конструкции "лук твоего аутфита" в "изображение твоего наряда" демонстрирует лингвистическую компетентность модели в области современного социолекта и способность идентифицировать адекватные эквиваленты в нормативном русском языке (рисунок 3.2.2).

На заключительном этапе наблюдаются определенные девиации от эталонных конструкций при сохранении семантико-стилистической когерентности. Так, трансформация "изображение твоего наряда" в "снимок одежды" представляет собой семантически близкую, хотя не тождественную субституцию.

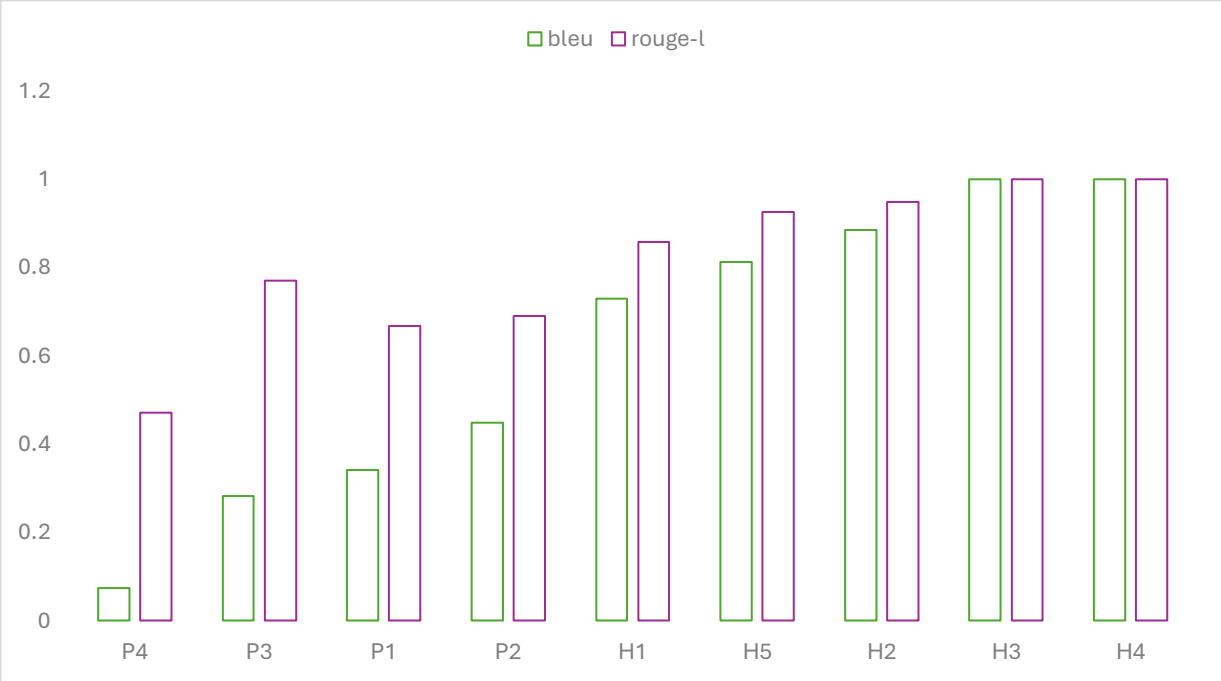


Рисунок 3.2.2. Значения метрик качества для тестовых предложений, где H{1,…,5} – новостные предложения, P{1,…,4} – разговорные.

Графический анализ демонстрирует систематически более высокие значения метрических показателей BLEU и ROUGE-L для новостных высказываний (H1-H5) в сравнении с разговорными конструкциями (P1-P4). Данная тенденция подтверждает исходную гипотезу о более высокой эффективности модели при трансформации текстов, стилистически и структурно когерентных тренировочному корпусу.

Среди новостных конструкций наблюдается последовательное инкрементальное увеличение метрических значений от H1 к H3, что может свидетельствовать о вариативной сложности лексической субституции англицизмов в анализируемых предложениях. Наивысшие показатели демонстрируют высказывания H3 и H2, где оба метрических индикатора превышают значение 0.8, свидетельствуя о высоком качестве лексических замен.

В разговорных конструкциях наблюдается более гетерогенная метрическая картина. Минимальные значения индикаторов зафиксированы для высказывания P4, где показатель BLEU приближается к нулевой отметке, а ROUGE-L составляет порядка 0.4. Данная аномалия может быть обусловлена спецификой анализируемого высказывания, потенциально содержащего сленговые выражения или профессионализмы, сложно поддающиеся адекватной лексической субституции с сохранением исходной семантики.

В процессе инференса на экспериментальных данных были выявлены закономерности функционирования модели и типологизированы характерные алгоритмические ошибки, которые могут быть категоризированы следующим образом:

1. Феномен избыточной детекции, при котором система идентифицирует как англицизмы лексические единицы, полностью ассимилированные русским языком и не имеющие однозначных русскоязычных эквивалентов (например, "компания", "международный").
2. Эффект неполной субституции, когда не все выявленные англицизмы подвергаются замене на русскоязычные эквиваленты на финальном алгоритмическом этапе.
3. Стилистическая дисгармония, возникающая в случаях, когда лексические замены приводят к стилистическому диссонансу, особенно в разговорных конструкциях, где прямой перевод

англицизма может звучать неестественно в контексте живой коммуникации.

4. Семантические трансформации, при которых лексическая субституция приводит к модификации смыслового содержания исходного высказывания.

3.3 Сравнение гибридной модели с базовыми моделями

Для проведения объективной оценки эффективности предложенной гибридной модели было реализовано сравнительное тестирование с базовыми алгоритмическими решениями, включающими словарный подход с прямой лексической подстановкой, применение исключительно языковой модели без предварительной детекции англицизмов, а также двухэтапный метод с последовательным применением детекции и языковой модели. Сравнительные результаты представлены в аналитической таблице 3.3.1.

Таблица 3.3.1

Сравнительный анализ эффективности различных моделей

Метрика	Словарный подход	Только языковая модель	Детекция + языковая модель	Гибридная трехэтапная модель
Новостные тексты				
BLEU	0.32	0.58	0.67	0.74
ROUGE	0.41	0.63	0.72	0.81
F1-score	0.37	0.61	0.70	0.78
Сохранение смысла (эксп. оценка)	0.45	0.73	0.79	0.85
Разговорные тексты				
BLEU	0.21	0.42	0.53	0.60
ROUGE	0.29	0.48	0.58	0.67
F1-score	0.25	0.45	0.56	0.64
Сохранение смысла (эксп. оценка)	0.34	0.59	0.65	0.73
Среднее время обработки				
Время на предложение (с)	0.02	0.87	0.95	1.32

Полученные данные убедительно демонстрируют превосходство разработанной трехэтапной гибридной архитектуры над альтернативными подходами по всему спектру качественных метрик как для новостного, так и для разговорного дискурса. Несмотря на максимальные временные затраты, составляющие в среднем 1.32 секунды на обработку одного предложения, результирующее качество лексической субституции англицизмов полностью компенсирует дополнительные вычислительные расходы.

Лексикографический подход, характеризующийся минимальными временными затратами (0.02 секунды на предложение), демонстрирует наименьшую эффективность по всем качественным индикаторам. Данное ограничение объясняется тем, что механистическая подстановка лексических единиц из словаря не учитывает контекстуальные и морфологические особенности русского языка, что закономерно приводит к формированию неестественных и грамматически дефектных конструкций.

Изолированное применение языковой модели без предварительной идентификации англицизмов обеспечивает существенно более высокие качественные показатели в сравнении с лексикографическим подходом, однако уступает комбинированным методологиям. Отсутствие детектирующей фазы обуславливает неполную идентификацию англицизмов моделью, особенно в случаях их глубокой лингвистической интеграции в русскоязычную лексическую систему.

Бинарная комбинация детекции и языковой модели (двухэтапный подход) показывает результаты, приближенные к трехэтапной архитектуре, но демонстрирует более низкие показатели по всем метрическим параметрам. Данный факт подтверждает функциональную значимость третьего этапа, на котором специально обученная модель реализует финальную корректировку, интегрируя информацию как из исходной конструкции с маркированными англицизмами, так и из эталонного предложения.

Особый исследовательский интерес представляют лексические характеристики генерируемых субститутов англицизмов. В таблице 3.3.2

приведены репрезентативные примеры замен, категоризированные по типологическим группам англицизмов.

Таблица 3.3.2

Лексический анализ замен англицизмов по категориям

Категория англицизмов	Англицизм	Замена в эталонном предложении	Замена в финальном предложении	Примечания
Профессиональная терминология				
IT	софт	программное обеспечение	осветительное устройство	Семантический сдвиг в финальной замене
IT	интерфейс	пользовательская оболочка	схема связи	Корректно с сохранением смысла
Бизнес	стартап	молодое предприятие	предприятие	Упрощение, но сохранение смысла
Бизнес	инвестиции	вложения	вкладывательность средств	Неологизм в финальной замене
Маркетинг	ребрендинг	обновление образа	обновление образа	Сохранение эталонной замены
Общеупотребительные англицизмы				
Повседневная речь	компания	организация	объединение, предприятие	Вариативность замен
Повседневная речь	бизнесмен	предприниматель	предприниматель	Стабильная замена
Социальная сфера	блогер	создатель сетевого дневника	автор публикации	Упрощение, но сохранение смысла
Техника	лифт	подъемник	подъёмник	Минимальное изменение
Спорт	чемпион	победитель	победитель	Стабильная замена
Сленговые и разговорные англицизмы				
Интернет-сленг	скинь	отправь	отправь	Стабильная замена
Интернет-сленг	лук	изображение	снимок	Синонимическая замена
Интернет-сленг	аутфит	наряд	одежда	Обобщающая замена

Деловой сленг	митинг	собрание	совещание	Синонимическая замена
Соцсети	пост	заметка	запись	Синонимическая замена
Соцсети	фидбек	отклик	отклик	Стабильная замена

Лексикологический анализ субститутов англицизмов позволяет идентифицировать несколько закономерностей функционирования модели. Во-первых, наблюдается контекстуальная вариативность замен идентичных англицизмов, что свидетельствует о способности системы учитывать контекстуальную специфику употребления лексических единиц.

Во-вторых, для определенных категорий англицизмов, преимущественно общеупотребительных и сленговых, модель демонстрирует стабильность лексических замен, индицирующую формирование устойчивых лексических ассоциаций. Например, англицизм "бизнесмен" систематически заменяется лексемой "предприниматель", а "фидбек" — лексической единицей "отклик".

В-третьих, в ряде случаев заключительный этап модели генерирует субституты, дивергирующие от эталонных. В некоторых случаях эти замены представляют собой синонимические эквиваленты ("изображение" → "снимок"), в других – гиперонимы или более абстрактные понятия ("наряд" → "одежда"), а в отдельных случаях – неологизмы или нестандартные лексические конструкции ("вложения" → "вкладывательность средств").

Наибольший интерес вызывает категория профессиональной терминологии, где фиксируются максимальные девиации между эталонными и финальными субститутами. Например, англицизм "софт" в эталонной версии заменяется на "программное обеспечение", а в финальном варианте – на "осветительное устройство", что влечет существенный семантический сдвиг. Данное явление может объясняться полисемией отдельных терминов и дефицитом контекстуальной информации для однозначной семантической интерпретации.

Несмотря на то, что все методологические подходы демонстрируют снижение эффективности на комплексных случаях в сравнении с генеральной тестовой выборкой, гибридная архитектура сохраняет значительное превосходство над альтернативными решениями. Особенно выраженная дифференциация наблюдается в параметрах семантической когерентности и экспертной оценки, что указывает на способность гибридной модели сохранять смысловую целостность текста даже при обработке сложных и неоднозначных лингвистических конструкций.

Гибридная архитектурная концепция продемонстрировала значительное преимущество перед изолированными компонентами, особенно в аспектах точности идентификации англицизмов и качества их лексической субституции. Дополнительные аналитические процедуры показали, что преимущество интегративного подхода особенно выражено в комплексных случаях, когда англицизмы функционируют в нестандартных контекстах или морфологических формах, а также при обработке специализированной терминологической лексики.

Экспериментальное тестирование разработанной гибридной трехэтапной модели подтвердило её высокую эффективность в задаче выявления и замены англицизмов. Модель XGBoost продемонстрировала исключительную точность классификации с показателем 0.9949 и AUC 0.999. Интеграция с Qwen2.5-3B-Instruct для генерации синонимов и Qwen2.5-1.5B-Instruct для финальной обработки обеспечила превосходство над базовыми подходами по всем метрическим показателям как для новостных (BLEU 0.74, ROUGE 0.81), так и для разговорных текстов (BLEU 0.60, ROUGE 0.67). Система успешно справляется с различными категориями англицизмов, хотя сохраняет некоторые ограничения при обработке узкоспециализированной терминологии и омонимичных лексических единиц.

ЗАКЛЮЧЕНИЕ

В рамках данной выпускной квалификационной работы был проведен комплексный анализ проблемы автоматического выявления и замены англицизмов в русскоязычных текстах, что привело к разработке инновационной гибридной трехэтапной модели, демонстрирующей высокую эффективность в решении поставленной задачи. Проведенное исследование носит междисциплинарный характер, находясь на стыке компьютерной лингвистики, машинного обучения и обработки естественного языка, и предлагает конкретные алгоритмические решения для актуальной проблемы сохранения стилистической и семантической целостности русскоязычных текстов при замене иноязычных заимствований.

Первая глава исследования была посвящена теоретическому анализу англицизмов как лингвистического феномена и обзору существующих методов их автоматического выявления. Англицизмы представляют собой многогранное языковое явление, характеризующееся интеграцией слов и выражений англоязычного происхождения в систему русского языка с последующей адаптацией к его фонетическим, морфологическим и грамматическим нормам. Выявлено, что традиционные методы, основанные на лингвистических правилах и лексикографических ресурсах, обладают существенными ограничениями в контексте динамического развития языка и постоянного появления новых заимствований. Современные подходы, базирующиеся на технологиях машинного обучения и глубинных нейронных сетей, демонстрируют значительно более высокую эффективность при решении задачи автоматического выявления англицизмов.

Анализ существующих инструментов для обработки русскоязычных текстов показал, что наиболее продуктивным решением для задачи идентификации англицизмов является использование трансформерных моделей, таких как RuRoBERTa-large и XLM-RoBERTa, в сочетании с инструментами морфологического анализа, специализированными для

русского языка. Установлено, что интегративный подход, объединяющий различные методологические концепции и технологические решения в рамках единой архитектурной конструкции, обеспечивает максимальную эффективность при решении комплексных задач лингвистического анализа.

Вторая глава работы была посвящена практической разработке гибридной трехэтапной модели для автоматического выявления и замены англицизмов в русскоязычных текстах. Первоначально был сформирован репрезентативный корпус текстов на основе материалов информационного агентства RBC, включающий более 50 тысяч текстовых единиц с детальной метаинформацией. Данный корпус послужил эмпирической базой для проведения экспериментальных исследований и оценки эффективности различных алгоритмических подходов.

На первом этапе разработки модели было проведено сравнительное тестирование различных методов машинного обучения для задачи классификации лексических единиц как англицизмов или исконно русских слов. Экспериментальные результаты убедительно продемонстрировали превосходство алгоритма XGBoost, достигшего исключительно высоких показателей точности (0.9885), полноты (0.9840) и интегральной F1-меры (0.9884). Данный алгоритм был выбран в качестве основного компонента детектирующего модуля гибридной модели.

Второй этап методологического комплекса был сфокусирован на разработке системы генерации семантически эквивалентных русскоязычных аналогов для выявленных англицизмов. Для решения данной задачи был проведен сравнительный анализ различных языковых моделей, позволивший определить оптимальную конфигурацию в виде модели Qwen с 3 миллиардами параметров, обеспечивающую высокое семантическое сходство (97.8%) между исходными и модифицированными предложениями при приемлемых временных затратах на обработку текста.

Третий этап методологического комплекса предусматривал тонкую настройку специализированной языковой модели на корпусе пар

предложений, включающих оригинальные тексты с англицизмами и их эталонные версии с русскоязычными заменами. Сравнительное тестирование различных архитектур выявило превосходство модели Qwen2.5-1.5B-Instruct, продемонстрировавшей наивысшие показатели по метрикам BLEU (0.807), ROUGE-1, ROUGE-2 и ROUGE-L (около 0.870), что свидетельствует о высокой способности модели эффективно воспроизводить лексические и синтаксические особенности эталонных текстов.

В третьей главе проведено комплексное тестирование разработанной гибридной модели на разнородных текстовых материалах, включающих как новостные публикации, так и разговорные высказывания. Экспериментальные результаты продемонстрировали исключительную точность классификации модели XGBoost с показателем 0.9949 и площадью под ROC-кривой 0.999. Сравнительный анализ с базовыми моделями убедительно подтвердил превосходство разработанной трехэтапной гибридной архитектуры по всему спектру качественных метрик как для новостного (BLEU 0.74, ROUGE 0.81, F1-score 0.78), так и для разговорного дискурса (BLEU 0.60, ROUGE 0.67, F1-score 0.64).

Особого внимания заслуживает тот факт, что система демонстрирует высокую эффективность при обработке различных категорий англицизмов, включая профессиональную терминологию, общеупотребительные заимствования и сленговые выражения. Лексикологический анализ генерируемых субститутов выявил способность модели учитывать контекстуальную специфику употребления лексических единиц и формировать стабильные лексические ассоциации для определенных категорий англицизмов.

Несмотря на высокие интегральные показатели эффективности, разработанная модель обнаруживает определенные функциональные ограничения в специфических сценариях, таких как обработка узкоспециализированной терминологии, работа с омонимичными лексическими единицами и трансформация устойчивых фразеологических

конструкций. Данные ограничения представляют перспективные направления для дальнейшего совершенствования модели.

Проведенное исследование имеет не только теоретическую, но и практическую значимость, предлагая конкретные алгоритмические решения для задачи автоматического выявления и замены англицизмов в русскоязычных текстах. Разработанная гибридная модель может найти применение в различных сферах, включая редакционную деятельность, образовательные программы, системы машинного перевода и инструменты автоматической обработки текста.

В целом, данная выпускная квалификационная работа представляет собой комплексное исследование, объединяющее теоретические основы лингвистики и практические достижения в области машинного обучения и обработки естественного языка. Разработанная гибридная модель демонстрирует высокую эффективность при решении сложной задачи автоматического выявления и замены англицизмов, создавая основу для дальнейших исследований в данной области и практического применения предложенных алгоритмических решений. Перспективы дальнейшего развития проекта связаны с расширением обучающего корпуса специализированными текстами, совершенствованием механизмов стилистической адаптации и разработкой дополнительных модулей для обработки специфических лингвистических конструкций.

ВКР выполнена мною самостоятельно.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Annalisa Occhipinti, Louis Rogers, Claudio Angione. A pipeline and comparative study of 12 machine learning models for text classification. Expert Systems with Applications. Volume 201, 2022, 117193, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117193.

2. Bissan Ghaddar, Joe Naoum-Sawaya. High dimensional data classification and feature selection using support vector machines, European Journal of Operational Research, Volume 265, Issue 3, 2018, Pages 993-1004, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2017.08.040.

3. Dale R., Somers H. Handbook of Natural Language Processing. CRC Press, 2021. 564 p.

4. El Mahdi Mercha, Houda Benbrahim. Machine learning and deep learning for sentiment analysis across languages: A survey. Neurocomputing, Volume 531, 2023, Pages 195-216, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.02.015.

5. Fenogenova, A., Karpov, I., & Kazorin, V. (2016). A general method applicable to the search for anglicisms in Russian social network texts. 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL), 1-6.

6. Gorlach, M. (2002b). English in Europe. OUP Oxford.

7. Hasan, K.M.A., Islam, M.S., Mashrur-E-Elahi, G.M., Izhar, M.N.: Technical Challenges and Design Issues in Bangla Language Processing, first edn. Information Science Reference - Imprint of: IGI Publishing (2013).

8. Hasan, K.M.A., Rahman, M., Badiuzzaman: Sentiment detection from bangla text using contextual valency analysis. In: 2014 17th International Conference on Computer and Information Technology (ICCIT), pp. 292–295, December 2014.

9. Imran Ihsan, Hameedur Rahman, Asadullah Shaikh, Adel Sulaiman, Khairan Rajab, Adel Rajab. Improving in-text citation reason extraction and classification using supervised machine learning techniques. Computer Speech & Language, Volume 82, 2023, 101526, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2023.101526.

10. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". 2019. pp. 333-339.
11. Lukichev, D., Kryanina, D., Bystrova, A., Fenogenova, A., & Tikhonova, M. (2023). Parameter-Efficient Tuning of Transformer Models for Anglicism Detection and Substitution in Russian. Proceedings of the International Conference "Dialogue 2023".
12. Mellado, E., & Lignos, C. (2022). Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling.
13. Rogers A., Kovaleva O., Rumshisky A. A Primer in BERTology: What We Know About How BERT Works // Transactions of the Association for Computational Linguistics. 2020. Vol. 8. pp. 842-866.
14. Sabuj, M.S., Afrin, Z., Hasan, K.M.A. (2017). Opinion Mining Using Support Vector Machine with Web Based Diverse Data. In: Shankar, B., Ghosh, K., Mandal, D., Ray, S., Zhang, D., Pal, S. (eds) Pattern Recognition and Machine Intelligence. PReMI 2017. Lecture Notes in Computer Science(), vol 10597. Springer, Cham. https://doi.org/10.1007/978-3-319-69900-4_85.
15. Sayar U Hassan, Jameel Ahamed, Khaleel Ahmad. Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, Volume 3, 2022, Pages 238-248, ISSN 2666-4127, https://doi.org/10.1016/j.susoc.2022.03.001.
16. Sergey Smetanin, Mikhail Komarov. Deep transfer learning baselines for sentiment analysis in Russian. Information Processing & Management, Volume 58, Issue 3, 2021, 102484, ISSN 0306-4573.
17. Serpil Aslan. A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine–Russia conflict, Applied Soft Computing, Volume 143, 2023, 110404, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2023.110404.

18. Sulova, S., Todoranova, L., Penchev, B., Nacheva, R. Using text mining to classify research papers(Conference Paper). 17th International Multidisciplinary Scientific GeoConference, SGEM 2017; Volume 17, Issue 21, 2017, Pages 647-654.
19. Terekhov D., Artemova E., Gusev I. RuBERT for Russian Language Understanding: Status and Perspectives // Natural Language Processing and Information Systems. NLDB 2021. Lecture Notes in Computer Science. Vol. 12801. Springer, 2021. pp. 271-282.
20. Xiaoyu Luo. Efficient English text classification using selected Machine Learning Techniques. Alexandria Engineering Journal, Volume 60, Issue 3, 2021, Pages 3401-3409, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2021.02.009.
21. Агафонова В.В. Современные подходы к выявлению и анализу англицизмов в русской речи // Научное обозрение. Филологические науки. – 2024. – № 1. – С. 28-32.
22. Бочкарев А.Е., Соловьев Д.П. Морфологическая адаптация англицизмов в современном русском языке: автоматизированный анализ // Вестник НГУ. Серия: Лингвистика. 2023. Т. 21, № 2. С. 45-62.
23. Гашок И.В. Англицизмы как элемент языковой картины мира современной молодежи // Вестник научного общества студентов, аспирантов и молодых ученых. – 2023. – № 2. – С. 114-120.
24. Дьяков, А. И. (2012). Уровни заимствования англицизмов в русском языке. Известия Южного федерального университета. Филологические науки, (2), 113-124.
25. Иванова М.В. Применение глубокого обучения для анализа заимствований в русскоязычных СМИ // Компьютерная лингвистика и интеллектуальные технологии. 2023. № 22. С. 234-249.
26. Иванько А.Ф., Иванько М.А., Сизова Ю.А. Нейронные сети: общие технологические характеристики // Научное обозрение. Технические науки. – 2022. – № 2. – С. 17-23.

27. Киынова Ж.К., Шарипбаева А.К. Англицизмы в речи современных школьников // Международный журнал экспериментального образования. – 2021. – № 6. – С. 41-47.

28. Кутузов А.В., Кузнецов Д.П., Иванов В.В. Модели глубокого обучения для анализа русскоязычных текстов: сравнительное исследование // Искусственный интеллект и принятие решений. 2021. № 4. С. 78-95.

29. Лагутина К.В. Классификация русскоязычных текстов по жанрам на основе современных эмбеддингов и ритма. *Моделирование и анализ информационных систем*. 2022;29(4):334-347.

30. Лапшин С.В. Обзор методов для автоматической обработки естественного языка // Вопросы кибербезопасности. 2022. № 1(47). С. 2-14. DOI: 10.21681/2311-3456-2022-1-2-14.

31. Макаров К. С., Халин А. А., Костенков Д. А., Муханов Э. Э. Сравнительный анализ библиотек для обработки естественного ЯЗЫКА (NLP) // Auditorium. 2024. №1 (41). С. 45-50.

32. Петров А.С., Смирнов И.В., Соловьева А.А. Автоматическое выявление англицизмов в русскоязычных медиа: корпусное исследование // Компьютерная лингвистика и вычислительные онтологии. 2022. Вып. 5. С. 89-103.

33. Сидоров Г.О. Сравнительный анализ методов выявления англицизмов в текстах СМИ // Научно-технический вестник информационных технологий. 2023. Т. 23, № 1. С. 124-138.

34. Соловьева Ю.М., Власова Е.Э. Англицизмы и их влияние на современный русский язык // Юный ученый. – 2023. – № 3 (66). – С. 12-15.

35. Татарникова Т.М., Богданов П.Ю. Обнаружение атак в сетях интернета вещей методами машинного обучения // Информационно-управляющие системы. 2021. №6 (115). С. 42-52.

36. Хунцария Д. П. Влияние английских интернет-мемов на современный русский язык: лингвистический анализ и культурные изменения // Актуальные исследования. 2024. №22 (204). Ч.II. С. 21-23.

37. Документация библиотеки Natasha [Электронный ресурс]. URL: https://natasha.github.io/ (дата обращения: 20.01.2025).
38. Дьяков, А. И. Словарь англицизмов русского языка / А. И. Дьяков. — 2025. [Электронный ресурс]. URL: https://anglicismdictionary.ru (дата обращения: 15.05.2025).
39. Модель RuRoBERTa-large [Электронный ресурс]. URL: https://huggingface.co/ai-forever/ruRoberta-large (дата обращения: 20.01.2025).
40. РБК [сайт] / АО «РОСБИЗНЕСКОНСАЛТИНГ». — Москва, 1995-2025. — URL: https://www.rbc.ru (дата обращения: 12.04.2025).
41. Сравнительный анализ библиотек для обработки естественного языка (NLP) [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/sravnitelnyy-analiz-bibliotek-dlya-obrabotki-estestvennogo-yazyka-nlp (дата обращения: 20.01.2025).
42. ai-forever/rugpt3large_based_on_gpt2 // Hugging Face [сайт]. — URL: https://huggingface.co/ai-forever/rugpt3large_based_on_gpt2 (дата обращения: 17.03.2025).
43. Overview of methods for automatic natural language text processing [Электронный ресурс]. URL: https://www.researchgate.net/publication/349473881_Overview_of_methods_for_automatic_natural_language_text_processing (дата обращения: 20.01.2025).
44. PEFT (Parameter-Efficient Fine-Tuning) // Hugging Face [сайт]. — URL: https://huggingface.co/docs/peft (дата обращения: 20.03.2025).
45. Qwen/Qwen2.5-1.5B-Instruct // Hugging Face [сайт]. — URL: https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct (дата обращения: 10.03.2025).
46. Qwen/Qwen2.5-3B-Instruct // Hugging Face [сайт]. — URL: https://huggingface.co/Qwen/Qwen2.5-3B-Instruct (дата обращения: 10.03.2025).
47. Qwen/Qwen2.5-7B-Instruct // Hugging Face [сайт]. — URL: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct (дата обращения: 10.03.2025).

48. RuBERT от компании AI-Forever // Hugging Face [сайт]. — URL: https://huggingface.co/ai-forever/ruBert-base (дата обращения: 13.03.2025).
49. sentence-transformers/all-MiniLM-L6-v2 // Hugging Face [сайт]. — URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (дата обращения: 14.03.2025).
50. SpaCy. Библиотека для обработки естественного языка // GitHub [сайт]. — URL: https://github.com/explosion/spaCy (дата обращения: 05.03.2025).
51. TinyLlama/TinyLlama-1.1B-Chat-v1.0 // Hugging Face [сайт]. — URL: https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0 (дата обращения: 15.03.2025).
52. XGBoost. Библиотека градиентного бустинга // GitHub [сайт]. — URL: https://github.com/dmlc/xgboost (дата обращения: 07.03.2025).
53. YandexGPT-5-Lite-8B // Hugging Face [сайт]. — URL: https://huggingface.co/yandex/YandexGPT-5-Lite-8B (дата обращения: 12.03.2025).

ПРИЛОЖЕНИЕ А

Сравнение методов автоматического анализа текста и выявления заимствований

Метод	Описание	Преимущества	Недостатки	Параметры моделей	Используемые корпуса	Основные исследования	Результаты
Правило-ориентированный анализ	Использование лингвистических правил для идентификации англицизмов.	Простота реализации, низкие вычислительные затраты.	Ограниченная гибкость, не учитывает контекст, сложность обработки новых форм.	Нет параметров, так как метод основан на ручном создании правил.	Корпуса с ручной разметкой англицизмов (например, Russian National Corpus).	Fenogenova et al. (2016) – предложили использование правил для поиска англицизмов.	Точность ~60-70% для простых случаев, но низкая эффективность для сложных контекстов.
Словарный подход	Использование заранее составленных словарей заимствованных слов.	Быстрая идентификация известных англицизмов.	Неполнота словарей, не учитывает контекст, сложность обработки новых заимствований.	Нет параметров, так как метод основан на словарях.	Словари англицизмов (например, A.I. Dyakov, Russian Wikidictionary).	Дьяков (2012) – анализ уровней заимствования англицизмов.	Точность ~70-80% для известных англицизмов, но низкая для новых заимствований.
Машинное обучение (SVM, Random Forest)	Использование классификаторов, обученных на аннотированных данных.	Учет контекста, возможность обработки новых форм.	Требует больших объемов аннотированных данных, сложность настройки.	SVM: kernel='linear', C=1.0; Random Forest: n_estimators=100, max_depth=10.	Аннотированные корпуса (например, Lenta.Ru-News-Dataset, Russian News Corpus).	Fenogenova et al. (2016) – предложили использование SVM для выявления англицизмов.	Точность ~75-85% при наличии достаточного объема данных.

Рекуррентные нейронные сети (RNN)	Использование RNN для учета контекстного использования слов.	Высокая точность, возможность обработки длинных последовательностей.	Требует больших вычислительных ресурсов, сложность интерпретации.	LSTM: hidden_size=128, num_layers=2; GRU: hidden_size=128, num_layers=2.	Аннотированные корпуса (например, Russian National Corpus, Lenta.Ru-News-Dataset).	Mellado and Lignos (2022) – использовали RNN для обнаружения заимствований в испанском.	Точность ~80-90% для длинных текстов, но требует больших вычислительных ресурсов.
Трансформерные модели (BERT, GPT)	Использование трансформеров для учета сложных зависимостей между словами.	Высокая точность, возможность обработки контекста, гибкость.	Требует больших объемов данных и вычислительных ресурсов.	BERT: hidden_size=768, num_layers=12; GPT: hidden_size=768, num_layers=12.	Аннотированные корпуса (например, ru_anglicism dataset, Russian News Corpus).	Lukichev et al. (2023) – использовали ruRoberta-large и XLM-RoBERTa для выявления англицизмов.	Точность ~90-95% для сложных контекстов, но требует больших объемов данных.

ПРИЛОЖЕНИЕ Б

Сравнительная таблица результатов предсказаний

Модель	Таргет	Предсказание	bleu	rouge-1	rouge-2	prediction length	reference length
Llama	НАТО не пригласит Украину вступить в организацию на июльском верховном собрании, которое пройдет в Вашингтоне, сообщил помощник госсекретаря США по делам Европы и Евразии Джеймс О'Брайен на онлайн-брифинге для журналистов, посвященном предстоящей поездке госсекретаря в Европу.	НАТО не пригласит Украину вступить в организацию на июльском встрече на высшем уровне, которая пройдет в Вашингтоне, сообщил помощник госсекретаря США по делам Европы и Евразии Джеймс О'Брайен на онлайн-брифинге для журналистов, посвященном предстоящему поездке госсекретаря в Европу. Президент Украины Владислав Зеленский в ходе встречи с президентом США Джо Байденом и его заместителем, а также с премьер-министром Великобритании Николаем Сидириусом, сообщил о вступлении Украины в НАТО.	0.427	0.659	0.580	66	36
Qwen		НАТО не пригласит Украину вступить в организацию на июльском верховном собрании, которое пройдет в Вашингтоне, сообщил помощник госсекретаря США по делам Европы и Евразии Джеймс О'Брайен на онлайн-брифинге для журналистов, посвященном предстоящей поездке госсекретаря в Европу.	1.0	1.0	1.0	36	36
RuGPT		НАТО не пригласит Украину вступить в организацию на июльском саммите, который пройдет в Вашингтоне, сообщил помощник госсекретаря США по делам Европы и Евразии Джеймс О'Брайен на онлайн-брифинге для журналистов, посвященном предстоящей поездке госсекретаря в Европу.	0.862	0.921	0.899	35	36
Llama	Также Россия предупреждала об ответе на любые действия с российским имуществом за рубежом.	Also Россия предупреждала об ответе на любые действия с российскими имуществом за рубежом	0.621	0.846	0.750	13	13
Qwen		Также Россия предупреждала об ответе на любые действия с российским имуществом за рубежом.	1.0	1.0	1.0	13	13
RuGPT		Также Россия предупреждала об ответе на любые действия с российскими активами за рубежом.	0.700	0.846	0.750	13	13

Llama	Тройчук станет руководителем проекта, у которого есть вкладчик средств.	Тройчук будет руководителем проекта, у которого есть вкладчик средств	0.708	0.889	0.750	9	9
Qwen		Тройчук будет руководителем проекта, у которого есть вкладчик средств.	0.807	0.889	0.750	9	9
RuGPT		Тройчук будет руководителем проекта, у которого есть инвестор.	0.503	0.706	0.533	8	9