

Пояснительная записка

к итоговому проекту на тему:

**"Классификация текстов"**

Автор: Власов Антон Алексеевич

Ментор: Татьяна Булгакова

Группа: NLP-11

Оглавление

[1. Постановка задачи 3](#_Toc181424856)

[Описание исходной задачи работы 3](#_Toc181424857)

[Актуальность задачи 3](#_Toc181424858)

[Проблема 3](#_Toc181424859)

[Метрики качества 4](#_Toc181424860)

[Методика оценки 6](#_Toc181424861)

[2. Анализ данных 7](#_Toc181424862)

[Исходные данные: 7](#_Toc181424863)

[ПРЕДОБРАБОТКА ДАННЫХ 7](#_Toc181424864)

[ЛЕММАТИЗАЦИЯ 8](#_Toc181424865)

[Разделение на обучающую и тестовую выборки 9](#_Toc181424866)

[Анализ аналогичных решений 9](#_Toc181424867)

[3. Методика реализации 11](#_Toc181424868)

[Итоговая модель 13](#_Toc181424869)

[4. Итоги обучения модели 14](#_Toc181424870)

[5. Выводы 16](#_Toc181424871)

[6. Список источников 17](#_Toc181424872)

1. Постановка задачи

Описание исходной задачи работы

Исходной задачей работы является классификация описаний сайтов по одной из 56 категорий.

Актуальность задачи

Текстовая классификация является одной из ключевых задач в машинном обучении и обработке естественного языка. В современном мире объёмы текстовых данных растут экспоненциально, и эффективная обработка этой информации становится критически важной. Автоматическая классификация позволяет структурировать большие массивы данных, улучшая доступность информации для анализа и принятия решений.

Актуальность задачи обусловлена широким спектром приложений. В бизнесе это анализ отзывов клиентов, мониторинг упоминаний бренда, фильтрация спама и персонализация маркетинговых кампаний. В социальных сетях и онлайн-платформах автоматическая модерация контента обеспечивает безопасность пользователей. Кроме того, классификация улучшает алгоритмы поисковых систем и рекомендательных сервисов, предоставляя более релевантный контент.

С развитием методов машинного обучения, включая глубокое обучение и модели трансформеров, точность и эффективность систем классификации продолжают расти. Это открывает новые возможности для инноваций в различных сферах — от бизнеса до науки и социальной сферы. В условиях, когда информация играет ключевую роль, задача классификации текста остаётся крайне актуальной.

Проблема

Классификация текстов с использованием нейросетей представляет собой сложную задачу, связанную с несколькими проблемами, которые могут повлиять на ее эффективность и точность. Рассмотрим основные из них.

1. **Сложность обработки языка**

**Порядок слов:** В отличие от других типов данных, текстовые данные имеют значительное значение порядка слов. Например, фразы с одинаковыми словами могут иметь противоположные значения в зависимости от их расположения. Это создает трудности для нейросетей, которые требуют сохранения информации о последовательности слов [(2)](https://habr.com/ru/companies/meanotek/articles/256593/" \t "_blank)

**Многозначность и контекст:** Одно и то же слово может иметь разные значения в зависимости от контекста, что затрудняет его правильную интерпретацию нейросетью (4)

1. **Ограниченность обучающих данных**

**Недостаток данных:** Для достижения высокой точности классификации требуется большое количество обучающих примеров. Однако в реальных сценариях часто наблюдается нехватка данных, что ограничивает возможности обучения нейросетей [(3)](https://loginom.ru/blog/neural-classification" \t "_blank)

**Грязные данные:** Наличие шумов и нерелевантной информации в обучающих данных может привести к ухудшению качества классификации. Предобработка и очистка данных являются критически важными этапами [(3)](https://loginom.ru/blog/neural-classification" \t "_blank)

1. **Выбор признаков**

**Избыточность или недостаточность признаков:** Слишком большое количество признаков может усложнить модель и увеличить вычислительные затраты, тогда как недостаток признаков может привести к потере важной информации для разделения классов [(3)](https://loginom.ru/blog/neural-classification" \t "_blank)

**Качественная векторизация:** Традиционные методы векторизации, такие как "мешок слов", могут не учитывать порядок слов и семантические связи, что снижает точность классификации [(4)](https://education.yandex.ru/handbook/ml/article/nejroseti-dlya-raboty-s-posledovatelnostyami" \t "_blank)

1. **Архитектурные ограничения нейросетей**

**Сложность архитектуры:** Нейросети могут требовать сложных архитектур для эффективной работы с текстами, таких как рекуррентные или трансформерные сети. Однако настройка этих архитектур может быть сложной задачей и требует значительных вычислительных ресурсов (4,5)

**Проблемы с градиентами:** При обучении нейросетей могут возникать проблемы с затухающими или взрывающимися градиентами, что затрудняет обучение моделей на длинных последовательностях текста [(4)](https://education.yandex.ru/handbook/ml/article/nejroseti-dlya-raboty-s-posledovatelnostyami" \t "_blank)

1. **Интерпретируемость результатов**

**Непрозрачность моделей:** Нейросети часто рассматриваются как "черные ящики", что затрудняет понимание того, как они принимают решения. Это может быть проблемой в приложениях, где важна прозрачность и объяснимость результатов [(3)](https://loginom.ru/blog/neural-classification" \t "_blank)

Эти проблемы подчеркивают необходимость тщательного подхода к разработке систем классификации текстов на основе нейросетей, включая выбор правильных архитектур, методов предобработки данных и стратегий обучения для достижения наилучших результатов.

Метрики качества

**Accuracy**

Одной из наиболее простых, а поэтому и распространенной метрикой является точность. Она показывает количество правильно проставленных меток класса (истинно положительных и истинно отрицательных) от общего количества данных и считается следующим образом

Accuracy   
Однако, эта простота является также и причиной, почему её часто критикуют и почему она может абсолютно не подойти под решаемую задачу.

**Precision**

Эта метрика показывает количество истинно положительных исходов из всего набора положительных меток и считается по следующей формуле:

precision   
Важность этой метрики определяется тем, насколько высока для рассматриваемой задачи «цена» ложно-положительного результата.

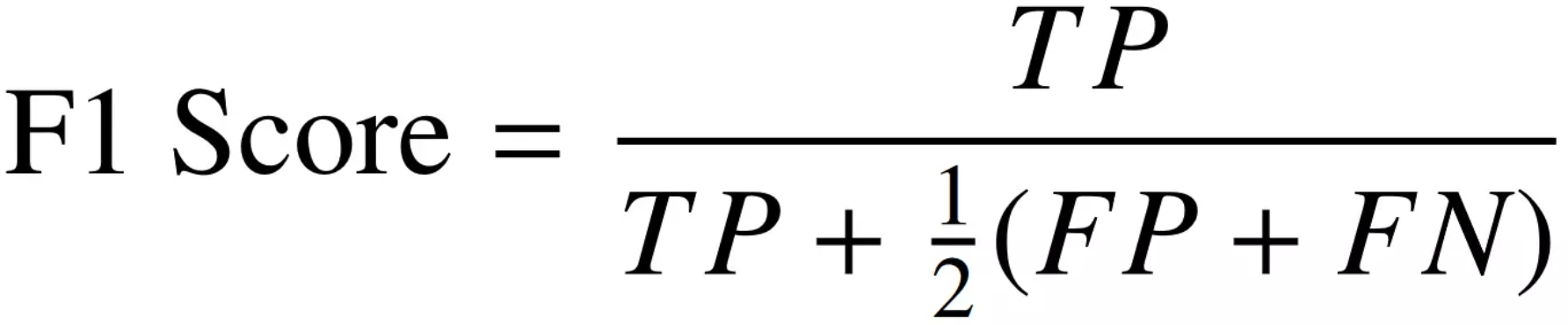
**Recall (true positive rate)**

На русском языке для этого термина используется слово «полнота» или «чувствительность». Эта метрика определяет количество истинно положительных среди всех меток класса, которые были определены как «положительный» и вычисляется по следующей формуле:

recall

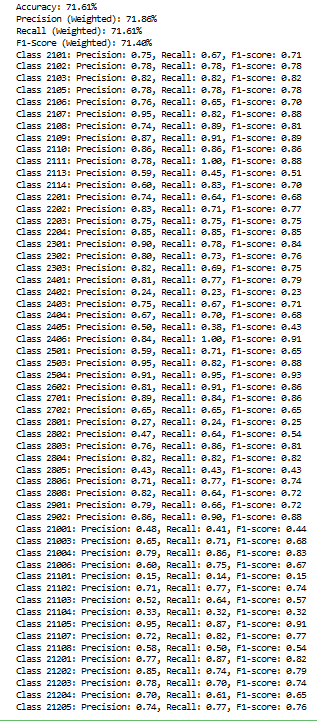
F1-Score

В том случае, если Precision и Recall являются одинаково значимыми, можно использовать их среднее гармоническое для получения оценки результатов:



Методика оценки

Для выведения метрик на экран использовались методы **accuracy\_score,** **precision\_score**  и **f1\_score** из библиотеки **scikit-learn**.



1. Анализ данных

Исходные данные:

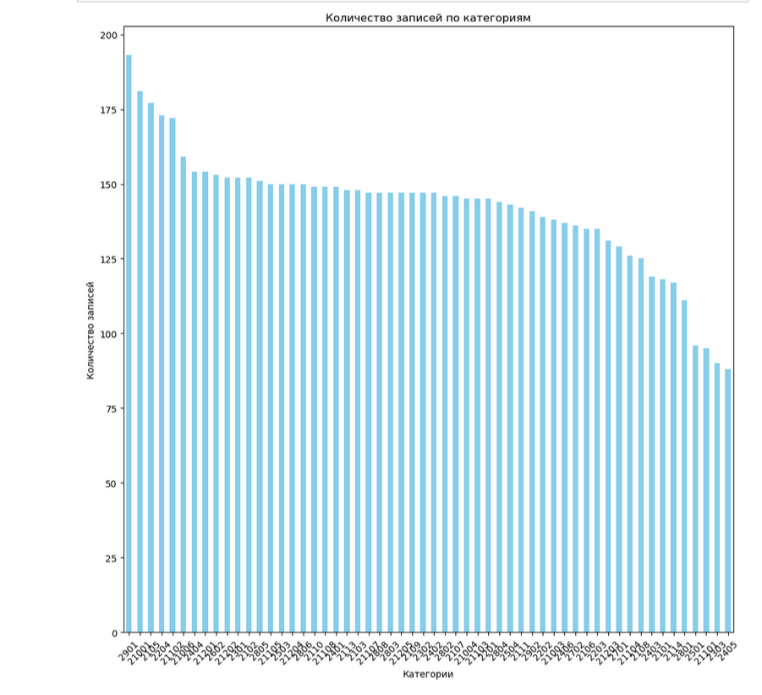
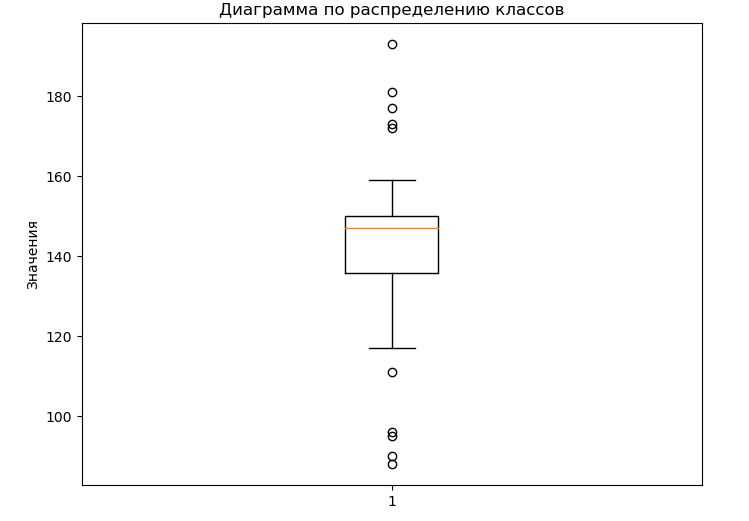
Датасет – приватный датасет с описаниями сайтов и соответствующим им числовым категориям

Количество текстов – 7951 шт

Количество категорий – 56.

Записи хранятся в CSV

Данные по категориям распределены равномерно (на каждую категорию приходится 1-2% от общего числа записей)



ПРЕДОБРАБОТКА ДАННЫХ

В ходе предобработки данных были проделаны следующие шаги:

* Удаление стоп-слов из библиотеки **spaCy** : ru\_core\_news\_sm
* Удаление предлогов и союзов
* Удаление знаков препинаний и непонятных символов
* Объединение нескольких пробелов в один

ЛЕММАТИЗАЦИЯ

Для лемматизации текстов использовалась библиотека **NLTK**

**Результат лемматизации**

ИСХОДНЫЙ ТЕКСТ

казино navi делает время казино более приятным. сделайте свое время в казино более приятным. начните с руководства казино casinonavi. изучайте веревки онлайн-казино, карточные игры блэкджек. получите ответы на все ваши вопросы, чтобы узнать, как выиграть в онлайн-казино. начни сейчас. о главных преимуществах использования биткойна в онлайн-казино. почему можно играть в онлайн-казино? знаете ли вы, сколько преимуществ игроки получают от онлайн-казино? лучшие советы, чтобы играть в онлайн-казино для начинающих. на самом большом веб-слоте играйте в pgslot без перерыва каждый момент. как найти хороший сайт для крипто-азартных игр.

ЛЕММАТИЗИРОВАННЫЙ ТЕКСТ

казино navi делать время казино приятный сделать свой время казино приятный начать руководство казино casinonavi изучать верёвка онлайн казино карточный игра блэкджек получить ответ весь ваш вопрос узнать выиграть онлайн казино начать главный преимущество использование биткойный онлайн казино почему играть онлайн казино знать сколько преимущество игрок получать онлайн казино хороший совет играть онлайн казино начинать сам большой веб слот играть pgslot перерыв каждый момент найти хороший сайт криптый азартный игра

ИСХОДНЫЙ ТЕКСТ

внчс проблемы решения главная страница. внчс проблемы решения. методики, по которым лечат внчс. полезные статьи внчс. как лечат внчс за ежом. re посоветуйте врача лфк. артроз, остеоартроз как ними бороться. примочки, спасающие от боли. философские, этические иные аспекты медицины.

ЛЕММАТИЗИРОВАННЫЙ ТЕКСТ

внчс проблема решение главный страница внчс проблема решение методика который лечить внчс полезный статья внчс лечить внчс ёж re посоветовать врач лфк артроз остеоартроз они бороться примочка спасать боль философский этический иной аспект медицина

ИСХОДНЫЙ ТЕКСТ

управление образования андроповского района. мероприятия, конкурсы , фестивали, соревнования. воспитательная работа и дополнительное образование. творческие конкурсы и мероприятия. профилактика детского дорожно транспортного травматизма. профилактика употребления наркотических и психоактивных веществ. дадим шанс детям. защита жилищных прав детей. независимая оценка качества образовательной деятельности. комплектование библиотечного фонда. документы по проверкам образовательных организаций. муниципальная система оценки качества образования. реализация майских указов президента россии. телефоны горячей линии по вопросам заработной платы. единое расписание егэ, огэ, гвэ. горячая линия гиа. есть предложения по организации учебного процесса или знаете, как сделать школу лучше. всероссийская олимпиада школьников. телефоны горячей линии по вопросам приёма граждан на обучение по образовательным программам.

ЛЕММАТИЗИРОВАННЫЙ ТЕКСТ

управление образование андроповский район мероприятие конкурс фестиваль соревнование воспитательный работа дополнительный образование творческий конкурс мероприятие профилактика детский дорожный транспортный травматизм профилактика употребление наркотический психоактивный вещество дать шанс ребёнок защита жилищный право ребёнок независимый оценка качество образовательный деятельность комплектование библиотечный фонд документ проверка образовательный организация муниципальный система оценка качество образование реализация майский указ президент россия телефон горячий линия вопрос заработный плата единый расписание егэ огэ гвэ горячий линия гиа предложение организация учебный процесс знать сделать школа всероссийский олимпиада школьник телефон горячий линия вопрос приём гражданин обучение образовательный программа

Разделение на обучающую и тестовую выборки

Количество категорий примерно равномерно распределено. Разделим датасет на обучающую и тестовую выборку, используя параметр stratify у train\_test\_split.

Размер обучающей выборки - 6763

Размер тестовой выборки - 1194

Анализ аналогичных решений

Для классификации текстов существует множество методов и подходов,которые можно разделить на классические машинные методы и нейросетевые решения.Рассмотрим их:

Классические машинные методы

* Мешок слов (Bag Of Word) - Представляет текст как набор слов без учета порядка и контекста. Каждое слово кодируется как одна из признаков в векторе. Не учитывает семантическую близость и игнорирует котекст слов
* TF-IDF – улучшение BoW, учитывает частоту слов в документах
* N-граммы – улучшение BoW, где рассматриваются последовательности из нескольких слов (биграммы, триграммы)
* Машинные алгоритмы (SVM, Random Forest, Наивный Байес) – используют вышеприведенные алгоритмы для получения признаков, затем используется классическое машинное обучение
* Word2Vec / GloVe / FastText – Преобразует слова в эмбеддинги с учетом семантической близости.
* Нейросетевые решения
* Рекурретные нейросети (RNN) - Могут учитывать контекст и запоминать предыдущие состояния
* LSTM/ GRU - Улучшенные версии RNN, решающие проблему исчезающего градиента.
* CNN – используют сверточные фильтры на уровне n-грамм.
* Трансформеры (BERT, S-BERT, DistilBERT, GPT) – используют механизмы внимания, позволяющие учитывать положение слова в контексте
* Гибридные методы
* Контекстные эмбеддинги – дополнительные контекстные эмбеддинги в качестве признаков моделей машинного обучения

Существующие решения для многоклассовой классификации, найденные на Kaggle, используют Random Forest, SVM, имеют в своей структуре гибридную модель (например, на основе линейной или сверточной нейронной сети) и показывают достаточно низкие результаты, не удовлетворяющие нашей задаче.

По результатам анализа существующих решений в области классификации текстов, наиболее подходящим для данной задачи был выбран гибридный подход – мы будем использовать эмбеддинги дообученного DistilBERT в сочетании с классическими нейронными сетями и механизмом внимания. Данная архитектура хорошо подходит под размер нашего датасета и обладает приемлемой производительностью.

1. Методика реализации

Для начала, необходимо обработать исходный датасет, лемматизировав слова, удалив стоп-слова и знаки препинания. Затем приведем категории к числовым значения от 0 до n-1, где n –число категорий.

Скачаем DistilBERT и дообучим его на трех эпохах на всех данных.

Создадим эмбеддинги предложений из усредненных значений слов на основании дообученного DistilBERT размера 768. Разделим выборку векторов на обучающую и тестовую при помощи train\_test\_split.  
Создадим линейную нейронную сеть со скрытым слоем, предсказывающую одну из категорий. Так как задача у нас достаточно сложная (необходимо предсказывать попадание текста в одну из 56 категорий), то скрытый слой будет по размерности больше, чем входной вектор – 4096. Для улучшения качества классификации добавим attention-слой.

**DistilBERT**

* DistilBERT — это упрощённая и более быстрая версия BERT, созданная с использованием техники, называемой Distillation (дистилляция знаний). Она позволяет обучать более компактные модели, сохраняя при этом большую часть производительности оригинальной модели. Основные различия между BERT и DistilBERT заключаются в следующем:
* Размер модели: DistilBERT примерно в два раза меньше, чем оригинальный BERT. Он использует на 40% меньше параметров, что упрощает хранение и развертывание модели.
* Производительность: Несмотря на уменьшение размера, DistilBERT достигает 97% точности BERT на большинстве задач, что делает его очень эффективным вариантом для использования в приложениях, требующих высоких скоростей обработки, например, в приложениях на мобильных устройствах.
* Обучение с дистилляцией: DistilBERT обучен так, чтобы "подражать" поведению более крупной модели (BERT). Во время обучения он использует результаты от BERT в качестве "подсказок", что помогает сохранить высокую точность при уменьшенном размере.
* Меньшее количество слоев: DistilBERT имеет 6 слоев трансформера вместо 12 в BERT-base, что ускоряет обработку данных и делает модель более экономичной в плане вычислительных ресурсов.
* Скорость: DistilBERT работает примерно на 60% быстрее, чем BERT, что существенно снижает задержки в реальных приложениях, таких как чат-боты и системы рекомендаций.

Мы будем использовать функцию активации ReLU

**ReLU (Rectified Linear Unit)**

ReLU (Rectified Linear Unit) - это нелинейная функция активации, которая широко используется

в глубоком обучении. Она преобразует входное значение в значение от 0 до положительной

бесконечности. Если входное значение меньше или равно нулю, то ReLU выдает ноль, в

противном случае - входное значение.

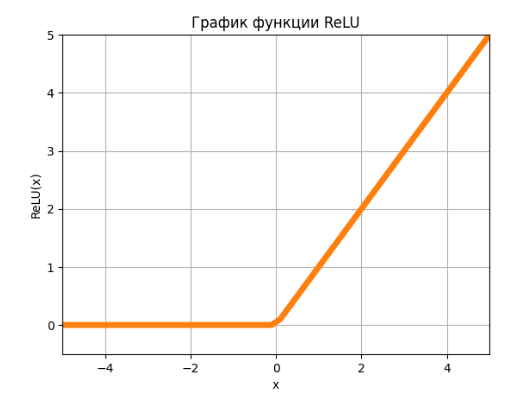
Математически ReLU определяется следующим образом:

ReLU(x) = max(0, x)

где max - функция, возвращающая максимальное значение из двух.

Графически ReLU выглядит как линейная функция с нулевым отсечением на оси абсцисс в точке

0. Это значит, что функция имеет постоянный наклон во всех точках, кроме точки 0, где

происходит отсечение.

ReLU имеет несколько преимуществ по сравнению со сигмоидной функцией активации. Во-

первых, ReLU более вычислительно эффективна, поскольку она является простой и быстрой

операцией, которая не требует вычисления экспоненты. Во-вторых, ReLU решает проблему

затухания градиента, так как она не вызывает затухания градиента при обратном

распространении ошибки, как это происходит в случае с сигмоидной функцией активации.

Однако, ReLU имеет некоторые недостатки. Во-первых, при использовании ReLU, некоторые

нейроны могут "умереть" (dead neurons), т.е. они могут получить отрицательное значение и

оставаться неактивными на всем протяжении обучения. Во-вторых, ReLU несимметрична

относительно нуля, поэтому может возникнуть проблема "расслоения" (clustering), когда

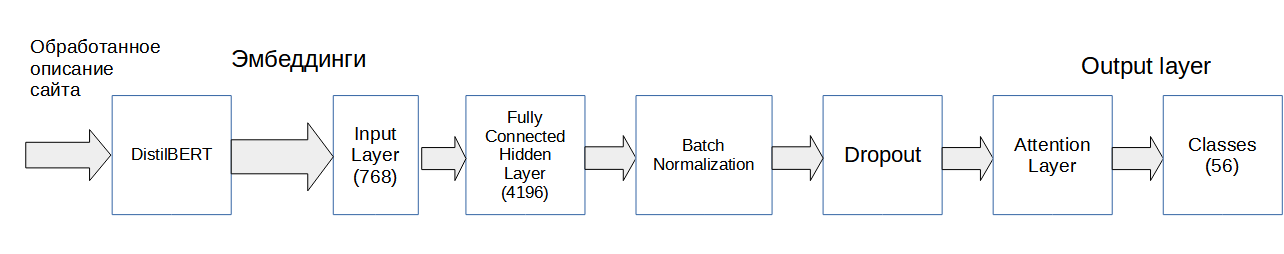
нейроны могут выдавать только положительные значения.

Итоговая модель

В итоговой модели используется оптимизатор Adam с начальной скоростью обучения 0,0001. Расчет ведется батчами по 4096 векторов. Используем DataLoader для упрощения обработки данных:

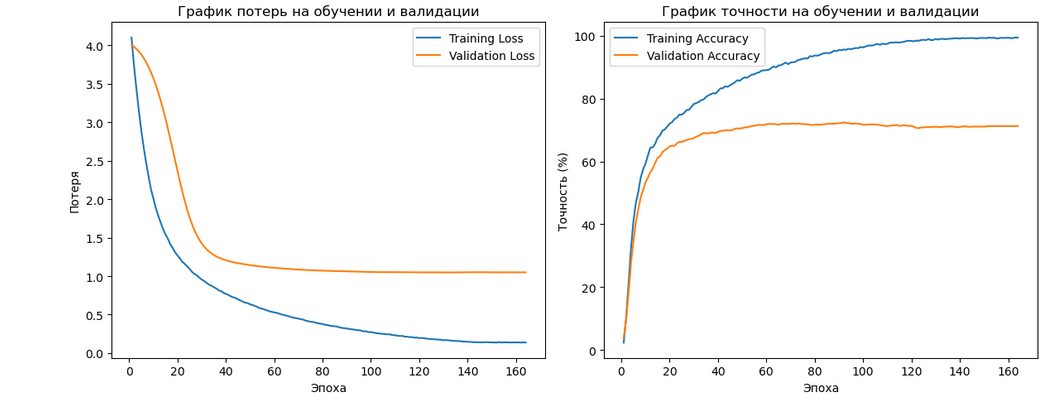
Дообучанием DistilBERT на протяжении 3 эпох на основе всех данных, таким образом увеличивая итоговую точность.

Для увеличения добавлена нормализация батчей и Dropout, а также слой внимания, расположенный перед нейронами, распознающими категорию.

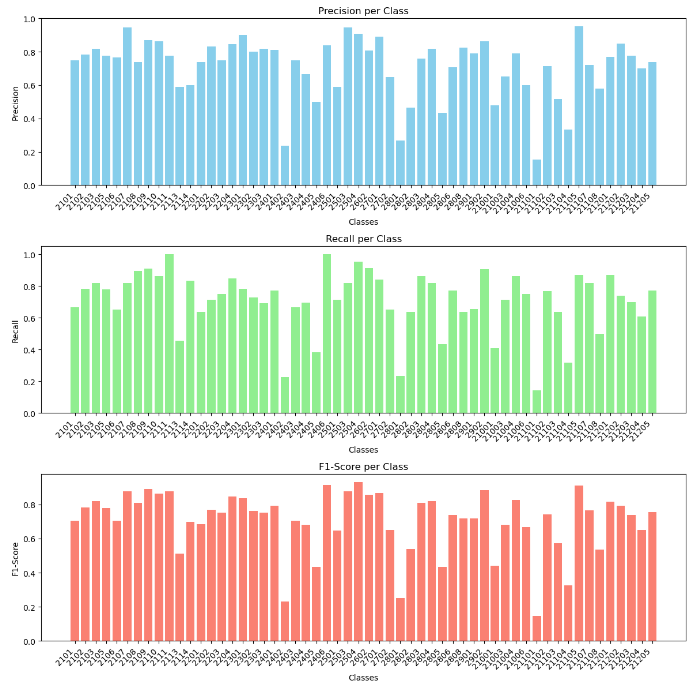


1. Итоги обучения модели

По результатам обучения 157 эпох (ранняя остановка, так как loss валидационной выборки в какой-то момент перестал уменьшаться), лучшая модель достигла F1 Score = 0.7140, Точности 0.7161 , Precision = 0.7186 , Recall = 0.7161



По итогам можно сказать, что модель научилась распознавать большую часть классов (см графики), однако, требует дальнейшего улучшения. Возможные варианты решения данной проблемы предложены далее.



1. Выводы

ходе проведенной работы была получена модель для категоризации описаний сайтов. Данная модель основана на эмбеддингах word2vec, линейной нейронной сети и слоя внимания. Целевые метрики составили 0.40-0.42

Для дальнейшей настройки модели рекомендуется:

* Увеличить количество данных по категориям, распознавание которых было выполнено неверно, или с небольшим значением метрик ( в частности, 2402, 2801 и 21101)
* Проводить кросс-валидацию
* Заняться более внимательным изучением тестовых данных, дополнительной обработкой, удалением большего числа незначащих слов
* Добавить аугментации в тексты для улучшения распознавания
* Провести доообучение DistilBERT.

целом, гибридная архитектура дает приемлемый результат для определенного набора входных данных при небольших вычислительных затратах, однако улучшение метрик требует более внимательного подхода к входным данным, требующий глубоких знаний в области работы с текстом и машинного обучения.

1. Список источников
2. <https://webiomed.ru/blog/osnovnye-metriki-zadach-klassifikatsii-v-mashinnom-obuchenii/>
3. <https://habr.com/ru/companies/meanotek/articles/256593/>
4. <https://loginom.ru/blog/neural-classification>
5. <https://education.yandex.ru/handbook/ml/article/nejroseti-dlya-raboty-s-posledovatelnostyami>
6. <https://ru.stackoverflow.com/questions/1520131/%D0%9D%D0%B5-%D0%BF%D0%BE%D0%BB%D1%83%D1%87%D0%B0%D0%B5%D1%82%D1%81%D1%8F-%D0%B4%D0%BE%D0%B1%D0%B8%D1%82%D1%8C%D1%81%D1%8F-%D1%82%D0%BE%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8-%D0%BE%D1%82-%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D1%81%D0%B5%D1%82%D0%B8-%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%82%D0%BE%D1%80%D0%B0-%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2>
7. Chatgpt.com
8. Huggingface.co
9. Perplexity.com