# Тема ВКР: Определение авторства текстов

Автор: Татьяна Фофанова
Научный руководитель: Сергей Аксенов

## Постановка и описание задачи

Задача определения авторства решается применительно к текстам русских, советских и российских писателей XVIII – XXI веков. Это задача многоклассовой классификации, с непересекающимися классами.

Набор данных содержит произведения 38 писателей — повести, романы, пьесы, рассказы, публицистику, воспоминания, письма и т.п. Поэтические произведения в данной работе не рассматриваются.

Основная задача, решаемая в работе — задача классификации текстовых данных; кроме основной, в работе предполагается решить ряд сопутствующих задач — кластеризация авторов по стилю, выделение семантической составляющей текста, генерация текста в заданном стиле.

Ожидаемые результаты:

- построение модели с использованием архитектуры трансформеров для классификации;
- интерпретация предсказаний классификатора и визуализация вклада токенов в предсказание;
- построение векторов эмбеддингов для писателей, описывающих их стиль;
- построение векторов эмбеддингов текстов, описывающих их семантику.

## Данные

В качестве исходных данных собраны произведения 38 авторов в EPUB-формате, список авторов и их произведений приводится в Приложении 1.

Данные для обучения модели представляют собой отрывки произведений длиной не менее 2000 символов, состоящие из целых предложений.

В обучающую выборку вошли 178 томов и отдельных произведений, по ним были получены более 47 тыс. объектов; в тестовой выборке – 79 томов / произведений, более 18 тыс. объектов.

Обучающая и тестовая выборки не пересекаются по произведениям, т.е. все отрывки из каждого произведения находятся либо в обучающей, либо в тестовой выборке. Это позволит избежать завышения метрик качества из-за неправильного дизайна эксперимента.

Количество и объем произведений распределены по авторам неравномерно, наблюдается несбалансированность по классам. Для приведения данных к более сбалансированному виду предполагается увеличить выборку отрывков для малых классов путем добавления отрывков с перекрытием. Т.е. отрывки для таких классов будут частично совпадать друг с другом.

Данные размещены в виде датасета на kaggle, их можно скачать с помощью kaggle API:

*kaggle datasets download --unzip tatianafofanova/authorstexts*

# Разметка данных

Данные размечаются автоматически, поскольку директории с обучающими и тестовыми данными содержат по 38 поддиректорий, соответствующих авторам. Каждая из этих поддиректорий содержит файлы EPUB-книг автора.

# Предобработка текстов

Каждая книга представляет собой набор глав, в HTML или XML формате.

Книга может включать в себя вступление, комментарии, критику и т.п., написанные другими писателями — эти главы необходимо исключить из анализа, т.к. они относятся к другим классам (писателям) и испортят выборку текстов.

Кроме того, среди начальных параграфов главы может встречаться наименование произведение и фамилия автора. Поэтому начальные параграфы не используются для формирования отрывков текста.

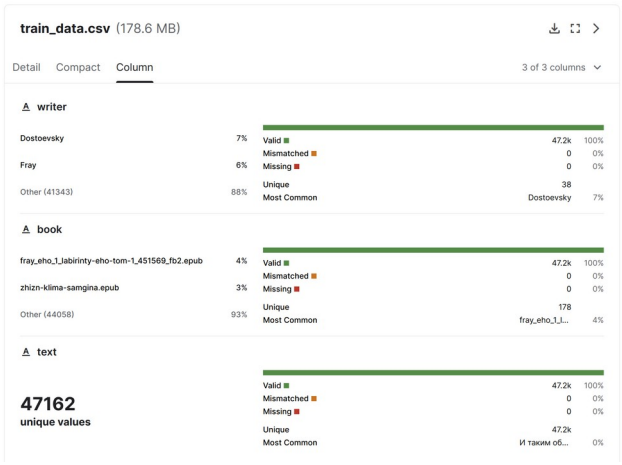
Все главы, в которых фамилия автора встречается более одного раза, рассматриваются как неавторские (написанные другими писателями), и игнорируются.

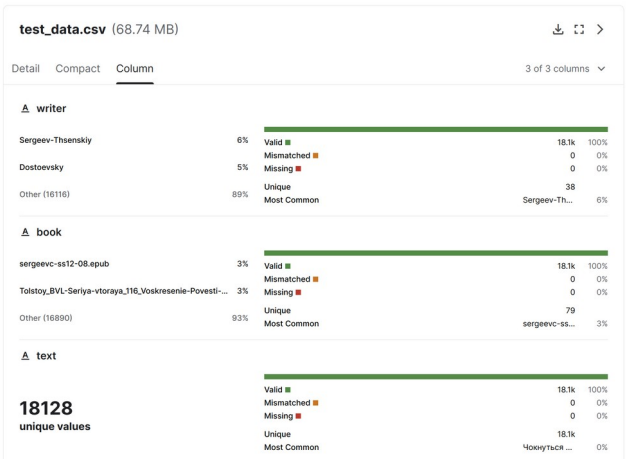
Из исходного текста главы удаляются все символы, кроме букв, цифр, знаков препинания, переноса строк и кавычек.

Затем из текста последовательно вырезаются отрывки длиной не менее 2000 символов, содержащие целые предложения.

# Пропущенные значения

В данных нет пропущенных значений, это подтверждается статистиками датасета на kaggle:



# Обзор литературы

Задача определения авторства текстов – задача многоклассовой классификации, которая может быть решена с помощью модели BERT. Существует множество доступных моделей BERT, которые были обучены на больших корпусах данных, с использованием больших вычислительных ресурсов. Для текущей задачи определения авторства текста использование подходящей предобученной модели позволяет начать обучение с хорошей стартовой точки, когда модель уже многое знает о структуре языка. Модель нужно дополнительно обучить, чтобы она научилась улавливать зависимости на корпусе литературных текстов и добавить слой классификатора для предсказания наиболее вероятного класса.

В статье [1] описаны методы, которые могут быть использованы для обучения классификатора на основе предобученной модели.

Рассматриваются следующие методы:

- использовать предобученную BERT для извлечения признаков для классификатора;

- дообучить предобученную BERT для решения задачи классификации, добавив в модель линейный слой, который получает на вход вектор токена [CLS] с последнего слоя энкодера, и затем применить функцию софтмакс для нахождения наиболее вероятного класса;

- дообучить маскированную линейную модель на произведениях русских / советских / российских писателей и затем использовать ее для решения задачи классификации как описано в предыдущем пункте. При этом можно использовать произведения и других писателей, не вошедших в датасет. Возможно даже использование переводных текстов, для дообучения языковой модели на неразмеченных данных.

В статье [2] предложен метод обучения эмбеддингов меток классов одновременно с обучением эмбеддингов токенов документа, это позволяет несколько улучшить качество классификации по сравнению с BERT.

В статье [5] описана модель BertGCN, совмещающая идеи обучения большой языковой модели и представления документов в виде графа, где вершинами являются документы и слова, входящие в них. В таком подходе предсказание метки класса опирается не только на собственно документ, но и на его соседей в графе. Такой подход дал заметное улучшение качества классификации.

Для интерпретации предсказаний модели, в статье [3] был предложен метод интегрированных градиентов, позволяющий оценить вклад каждого токена входной последовательности в предсказание класса. Воспользуемся этим методом для исследования предсказаний классификатора.

В статье [4] исследуется вопрос о способности модели BERT получать знания о структуре языка. Это полезно для понимания, какую информацию содержат скрытые состояния модели на разных слоях, и может быть использовано для построения векторов эмбеддингов авторов и эмбеддингов текстов, отражающих семантический смысл.

# Разведывательный анализ данных

Данные содержат метку автора, название файла (EPUB-книги), из которого взят отрывок, и текст отрывка.

В каждом файле может быть несколько произведений автора (сборник), одно произведение или часть произведения (том). В следствие этого посчитать количество произведений по автору затруднительно, и, кроме того, объектами датасета являются не произведения, а отрывки из них. Поэтому далее будем вычислять статистики для отрывков произведений.

Изучим, сколько текстов писателей и какой длины присутствуют в данных. Для этого для каждого автора вычислим количество текстов и медиану длины отрывков

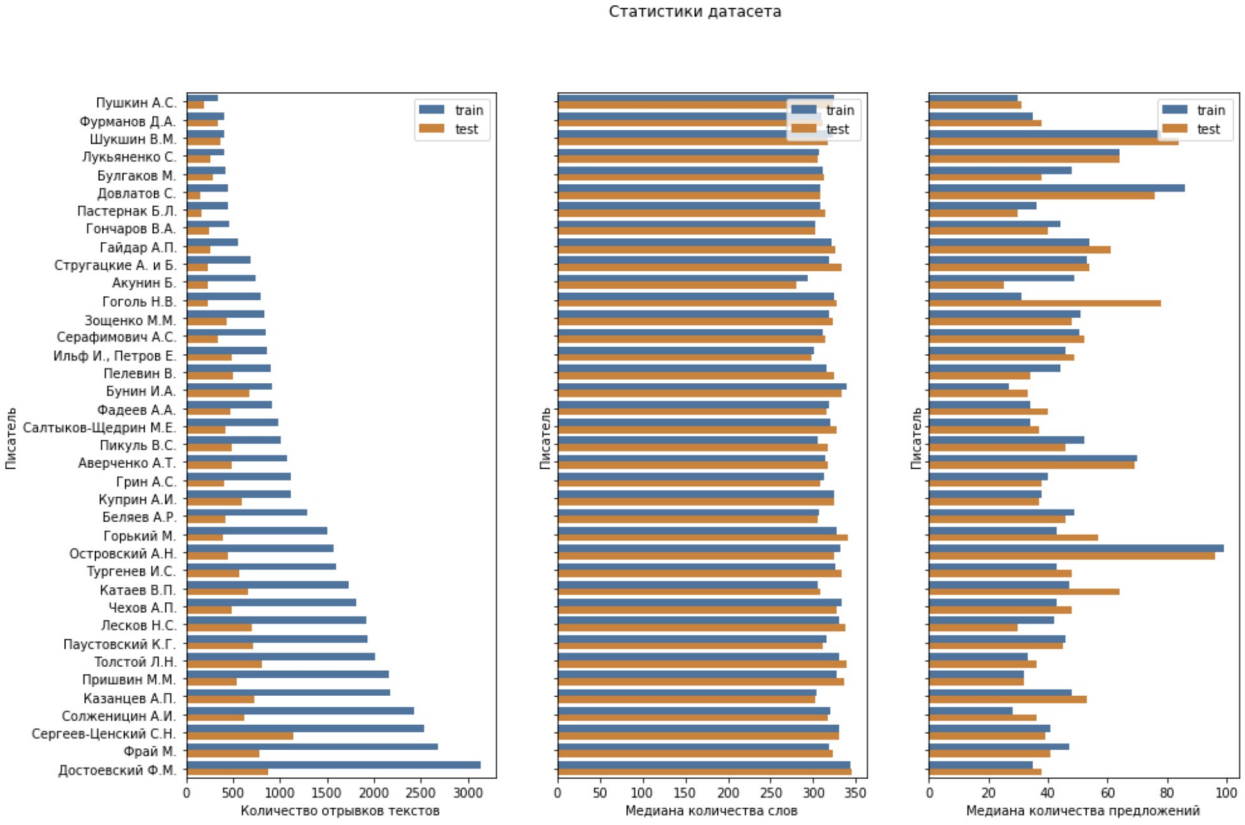



Рис.1. Характеристики обучающей и тестовой выборок.

Как было отмечено выше, данные не сбалансированы по классам – некоторые из классов малочисленные (Пушкин, Фурманов и т.п.), некоторые – наоборот, объемные (Достоевский, Солженицин).

Базовая модель будет построена на несбралансированом датасете, затем для сравнения будет построена аналогичная модель на сбалансированном по классам наборе данных. Небольшие по объему данных классы будут увеличены за счет добавления отрывков текстов с перекрытиями, а слишком большие классы будут урезаны с помощью случайного отбора некоторого фиксированного количества текстов для каждого автора.

На рис.1 также видно, что среднее количество слов в отрывках колеблется незначительно в диапазоне 300 – 350 слов, в то время как количество предложений в отрывке различается сильнее (от 20 до 90 предложений).

Большое количество предложений характерно для пьес, где каждая реплика сопровождается указанием действующего лица. В качестве предобработки текстов предполагается удалить из пьес наименование действующих лиц, оставив только реплики. Это может быть выполнено с помощью частотного анализа предложений, входящих в отрывок. Посмотрим на разброс частот предложений в обычной прозе и в пьесе (тексты отрывков приведены в Приложении 2):

Пьеса

| | frequency |
|---|---|
| Карп. | 12 |
| Аксюша. | 7 |
| Буланов. | 3 |
| А когда мне? | 1 |
| Боитесь… Чтоб были набиты! | 1 |
| Не скажете вы при них и курить-то боитесь. | 1 |
| Я вот Раисе Павловне скажу. | 1 |
| Вот что. | 1 |

Проза

| | frequency |
|---|---|
| Подумаешь, машина. | 2 |
| Сколько? | 1 |
| Может, она там? | 1 |
| Может, сойдет? | 1 |
| упавшим голосом выговорил писатель. | 1 |
| Вы не беспокойтесь. | 1 |
| У меня их штук двадцать пять. | 1 |
| Правда, у той был капот. | 1 |

Предложения с частотой более 2 могут быть удалены из отрывка, тогда пьесы и проза будут более похожи. Базовая модель будет обучена на текстах без такой предобработки, далее будет проведен эксперимент с предобработкой пьес.

В каждом отрывке текста могут присутствовать разные типы предложений — повествовательные, вопросительные, восклицательные. Вычислим для каждого писателя доли вопросительных, восклицательных предложений, предложений с многоточием и диалогов, а затем построим гистограммы для этих долей.

Как видим на рис.2., авторы с разной частотой используют вопросительные, восклицательные предложения, многоточие и диалоги. К примеру, большинство писателей редко используют диалоги, но для некоторых из них диалоги составляют 30% объема текста.

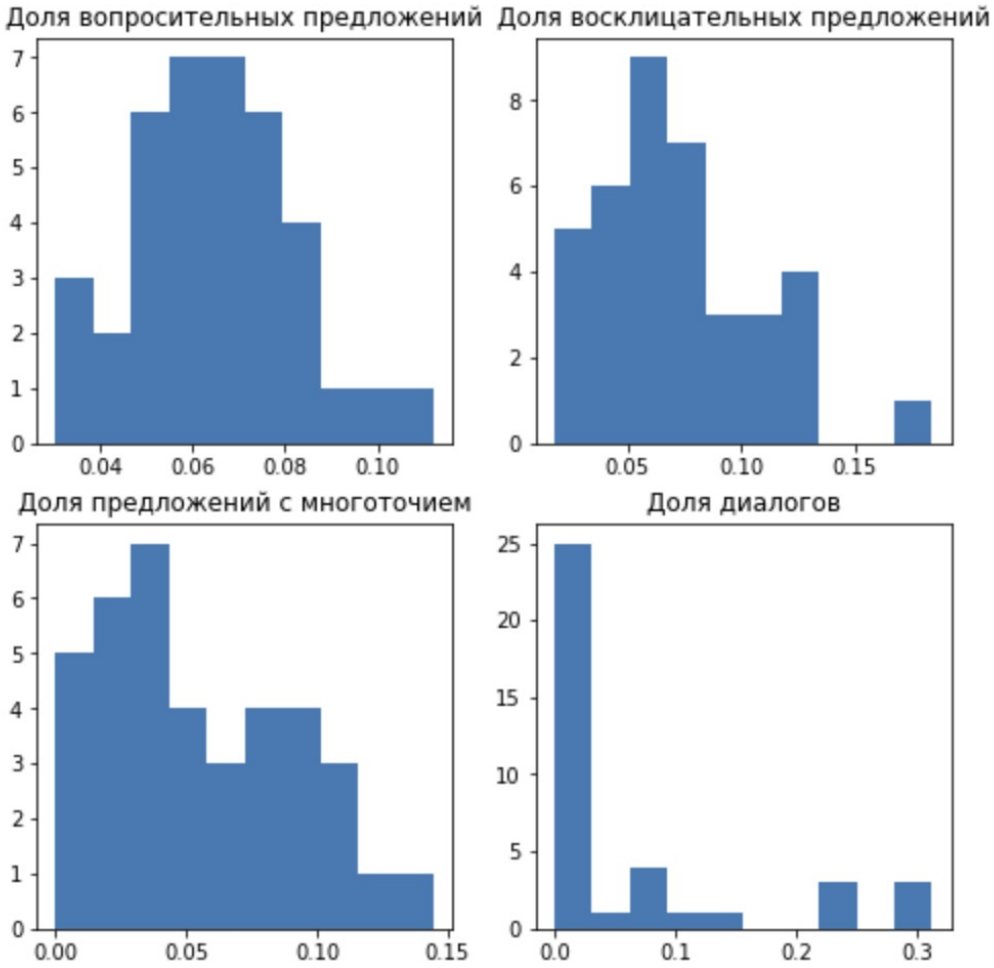



Рис. 2. Распределение долей предложений разного типа в текстах авторов.

# Построение базовой модели BERT

В качестве базовой модели классификации текстов была выбрана модель BERT, обученная на большом корпусе русскоязычных текстов, и дообученная на примерах из собранного датасета.

Архитектура модели BERT  построена на архитектуре Трансформер и включает в себя только слои кодировщика, примененные последовательно друг за другом.

Текстовые данные, поступающие в модель, преобразуются с помощью токенизера в последовательность токенов — слов / частей слов, из предобученного словаря. Далее токены заменяются на числовые индексы, чтобы данные, поступающие в модель имели числовое представление. На этом же этапе последовательности выравниваются по длине — обрезаются или заполняются служебными токенами [PAD], обозначающими нулевые значения. Начало и конец последовательности также заменяются служебными токенами [CLS] и [SEP].

На вход модель получает последовательность в виде индексов токенов, полученной от токенизера,  и бинарной маски, указывающей на то, какие из токенов являются входными данными, а не пустыми токенами [PAD].

Затем данные проходят через слои эмбеддингов и позиционных эмбеддингов, на выходе из которых для каждого токена получаем его векторное представление. После этого векторные представления проходят через 12 слоев кодировщика BERT.

Предобученная модель размещена на ресурсе huggingface и имеет следующие характеристики:

- путь: sberbank-ai/ruBert-base

- обучена на задаче предсказания маскированного слова

- архитектура: энкодер

- токенизер: BPE (byte pair encoding)

- размер словаря: 120 138

- количество параметров: 178 M

- объем обучающих данных: 30 Гб

Для обучения классификатора на базе BERT использовался класс AutoModelForSequenceClassification из библиотеки transformers, веса кодировщика которой инициализировались из предобученной модели, веса линейного слоя — случайными значениями.

Модель AutoModelForSequenceClassification представляет собой кодировщик BERT, состоящий из 12 слоев, и нескольких дополнительных слоев: слой dropout и линейный слой с 38 выходами, каждый из которых соответствует определенному автору. Линейный слой принимает на вход векторное представление первого токена из последнего слоя кодировщика

BERT — токена [CLS] из исходной последовательности. Векторное представление этого токена содержит в себе всю информацию о входящей последовательности, поэтому достаточно подать только его на линейный слой для дальнейшей классификации.

Обучение базовой модели проводилось в течение 5 эпох, с размером батча, равным 6.

После каждой эпохи вычислялись метрики качества модели. Лучшая модель определялась по метрике F1-score с макро-усреднением, чтобы вклад каждого класса в метрику был одинаковым и не зависел от размера класса.

Базовая модель имеет F1-score  с макроусреднением, равный 0.71, и с микроусреднением — 0.74.

Логирование эксперимента проводилось с помощью сервиса WandB, сохраненной в виде артефакта моделью можно воспользоваться следующим образом:

```
run = wandb.init()
artifact = run.use_artifact('sava_ml/Diploma/baseline38:v1', type='model')
artifact_dir = artifact.download()
```

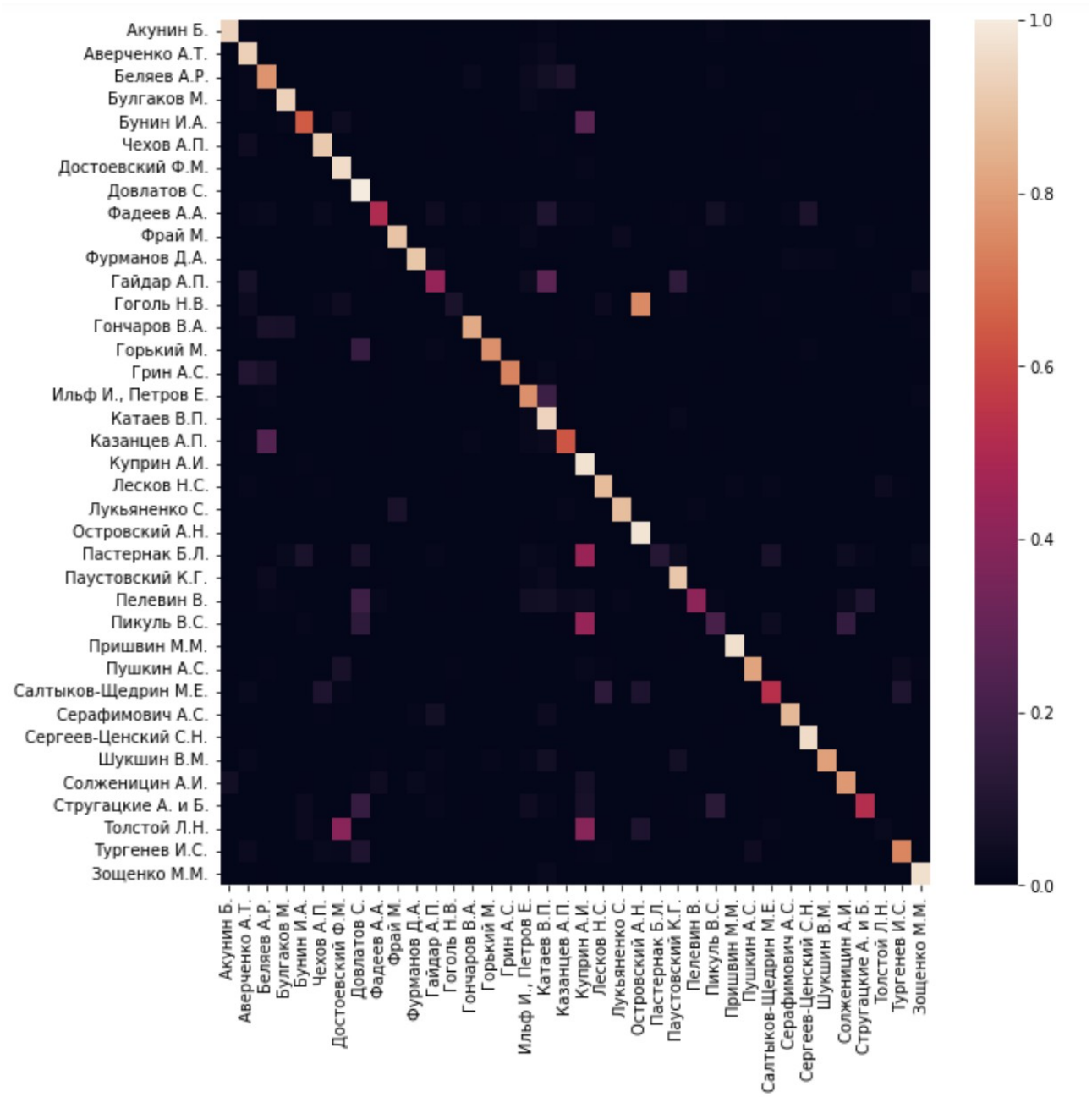
# Анализ ошибок классификатора

Для более глубокого понимания, в каких случаях модель делает наибольшее количество ошибок, посмотрим на значение f1-score для каждого класса в отдельности. В десять наиболее трудных для модели классов вошли:

| Автор | f1-score |
|---|---|
| Толстой Л.Н. | 0.02 |
| Гоголь Н.В. | 0.08 |
| Пастернак Б.Л. | 0.12 |
| Пикуль В.С. | 0.22 |
| Пелевин В. | 0.41 |
| Гайдар А.П. | 0.43 |
| Фадеев А.А. | 0.51 |
| Стругацкие А. и Б. | 0.52 |
| Салтыков-Щедрин М.Е. | 0.53 |
| Казанцев А.П. | 0.64 |

Видно, что Толстого Л.Н. и Гоголя Н.В. модель классифицирует неправильно в более чем 90% случаев - это очень низкий показатель качества.

Посмотрим на матрицу ошибок, чтобы выяснить, каких писателей модель часто путает между собой:




Заметим, что модель почти всегда путает Гоголя с Островским. Этому можно найти такое объяснение: в тестовые примеры для Гоголя вошли отрывки из произведений Ревизор и Женитьба. Драматические отрывки, оба произведения - пьесы. Поскольку Островский писал в основном пьесы, а среди обучающих данных Гоголя пьес не было, то модель могла переобучиться на формат произведения и все пьесы приписывать Островскому.

Кроме того, модель очень плохо классифицирует тексты Толстого Л.Н. Проанализировав разбиение на обучающие и тестовые тексты для этого автора, заметим, что в тестовые данные попали короткие произведения — рассказы и повести, в то время как в обучающие — в основном, большие романы.

Обнаруженная нерепрезентативность тестовых данных мешает и качественному обучению модели, и качественной оценке результатов обучения.

Наилучшим решением в такой ситуации является перебалансировка обучающей и тестовой выборки так, чтобы пьесы находились в обеих частях.

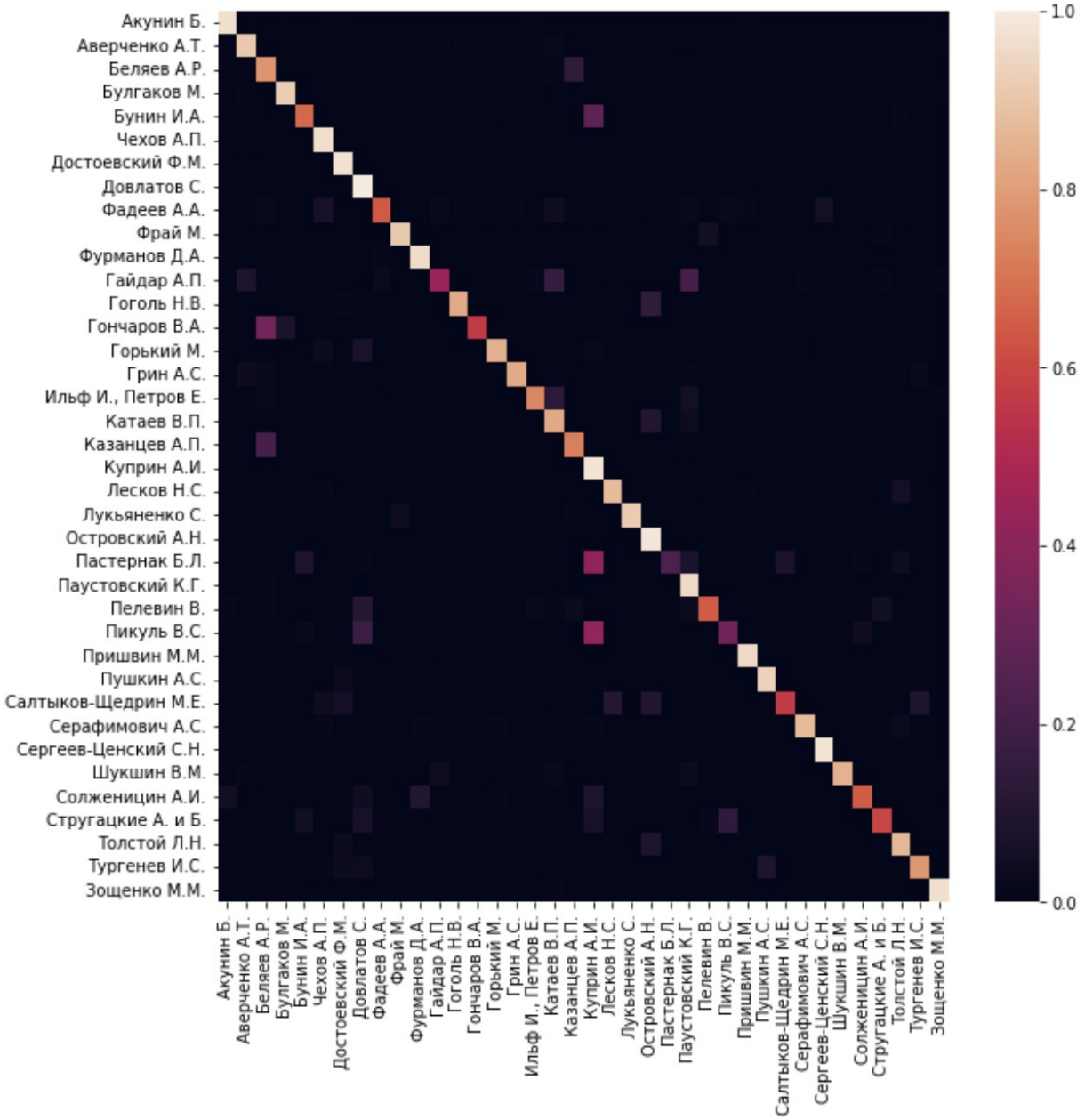
Кроме того, во многих отрывках текстов Толстого Л.Н. есть большие фрагменты на французском, и это может влиять на качество, т. к. токенизатор содержит только русские токены.

Для устранения этой проблемы добавим в предобработку удаление всех символов, кроме кириллицы и знаков препинания.

Качество классификатора существенно выросло, f1-score с макроусреднением достиг 0.795, с микроусреднением — 0.821.

Матрица ошибок после устранения недостатков первоначального разбиения на обучающую и тестовую выборки[1]:

1 Кроме перебалансировки данных, на данном этапе были внесены изменения в обучающую процедуру — размер батча был увеличен до 48, чтобы на каждом шаге оптимизации количество классов было меньше количества объектов, что делает обучение более стабильным.




На данном этапе были проведены эксперименты с дообучением BERT на собранных данных. Данный эксперимент также показал важность аккуратной работы с данными — некачественное разбиение на обучающую и тестовую выборки оказывает сильный негативный эффект на результирующую метрику.

Кроме этого, важно настраивать процесс обучения. В данном случае, накапливание батчей большего размера перед выполнением шага оптимизации позволило увеличить качество итоговой модели.

# Проведенные эксперименты, которые нужно повторить на исправленном датасете

## 1. Балансировка датасета

Поскольку классы в датасете несбалансированы, то модель обучается на разном количестве примеров для каждого класса. Это может привести к тому, что будет наблюдаться низкое качество классификации для малочисленных классов.

Для балансировки классов было увеличено количество отрывков для малочисленных классов так, что отрывки брались не последовательно, а имели перекрытие. Также было сокращено количество отрывков из многочисленных классов — максимальное количество отрывков для каждого автора было ограничено 1500 примерами.

Однако, при анализе ошибок классификации такой тенденции не обнаружилось, поэтому можно предположить, что балансировка датасета не улучшит качество базовой модели, что и подтвердилось опытным путем на исходных данных.

## 2. Дообучение языковой модели на неразмеченных текстах

Для улучшения качества предсказаний классификатора можно дообучить предобученную языковую модель на неразмеченных данных. На этом этапе языковая модель будет учиться на задаче предсказания маскированных токенов. После этого добавим к дообученной языковой модели классификатор и обучим их на имеющемся размеченном датасете.

В качестве неразмеченных данных, используемых для дообучения языковой модели, было использовано 3 корпуса данных:

- тексты корпуса Taiga proza_ru, содержащие тексты из литературных журналов 2005-2017 гг. (в основном, рассказы) на русском языке;

- отрывки литературных текстов из переводных произведений на русском языке;

- отрывки из имеющегося датасета.

Обучение языковой модели на корпусе Taiga не улучшило качество классификации, т. к. тексты, содержащиеся в этом датасете отличаются от текстов классических литературных текстов, применительно к которым решается исходная задача. В журналах печатаются, в основном, рассказы, тексты могут содержать сленг или обороты, характерные для современного разговорного языка. Такой подход не делает языковую модель более подходящей для решаемой задачи и снижает качество классификации.

Второй подход — дообучить языковую модель на фрагментах классических литературных произведений, переведенных на русский язык, кажется более перспективным.

## 3. Обучение модели SBERT

Модель SBERT представляет собой модификацию модели BERT, обученную на триплетной функции потерь. Такая функция будет стараться уменьшить расстояние между векторными

представлениями текстов, относящихся к одному классу, и увеличить расстояние между объектами разных классов.

Такой подход позволяет использовать BERT для нового типа задач, с которыми она не сталкивалась. Эти задачи включают в себя вычисление семантической близости последовательностей, кластеризация, поиск соответствия по семантическому поиску.

При использовании триплетной функцией потерь, модель получает на вход не один объект, а три — якорь, положительный пример (объект того же класса, что и якорь), и отрицательный пример (объект другого класса).

При обучении модели большую роль играет формирование таких троек — важно, чтобы модель училась на сложных для нее наборах, когда расстояние от якоря до отрицательного примера меньше, чем расстояние до положительного.

Получение качественных векторных представлений для текстов, близких к друг другу для одного класса, и далеких для разных, могло бы улучшить качество классификатора, построенного поверх таких эмбеддингов.

В работе было проведено два эксперимента:

- обучение модели SBERT и дальнейшее применение модели для получения признаков для классификатора. Параметры классификатора обучаются отдельно от SBERT;

- обучение модели SBERT и классификатора одновременно. При этом шаг оптимизации делается по каждой функции потерь — триплетная функция потерь для SBERT и кросс-энтропия для классификатора.

Для обучения модели SBERT использовался фреймворк sentence-transformers, позволяющий формировать тройки примеров для обучения, далее, учитывая сложность примеров, формировать батчи для обучения модели, подсчитывать функцию потерь.

В результате обучения SBERT, доля примеров с правильным соотношением расстояний составила 0.954.

С помощью метода понижения размерности признакового пространства t-SNE, построим распределение векторных представлений примеров из тестовой выборки.

На этих векторных представлениях будет обучен классификатор.

# Визуализация

Построенная модель предсказывает метку класса, и, чтобы ответить на вопрос, почему модель предсказала именно такую метку, нужно научиться интерпретировать предсказание модели.

В частности, можно вычислить вклад каждого элемента входящей последовательности, т. е. каждого токена, на итоговое предсказание класса. Для этого существует метод интегрированных градиентов.

Метод интегрированных градиентов позволяет оценить вклад каждого признака на результат, полученный на выходном слое модели. Этот метод может быть применен к любой модели глубинного обучения.

Для применения метода интегрированных градиентов, в модель подается две входных последовательности: исходная, оригинальная последовательность, для которой и проводится интерпретация предсказаний, и «нулевая»  последовательность, состоящая только из [PAD] токенов.

Затем «нулевая» последовательность постепенно, в течение m шагов, интерполируется в оригинальную последовательность. На 1-м шаге интерполяции последовательность будет мало отличаться от «нулевой», на m-1 — м шаге — последовательность будет почти оригинальной, на m-м шаге — полностью совпадать с оригинальной.

Влияние каждого признака входящей последовательности описывается формулой:

$$IG(\text{approx}) \approx (x_i - x_i^{'}) * \sum_{k=1}^{m} \frac{\delta F\left(x' + \frac{k}{m} * (x - x')\right)}{\delta x_i} * \frac{1}{m}$$

где i — индекс признака,

x — оригинальная входящая последовательность,

x“ - «нулевая» входящая последовательность,

k — номер шага интерполяции,

m — общее число шагов интерполяции.

Метод интеграционных градиентов реализован в библиотеке Captum.

Посмотрим, какие из токенов входящей последовательности оказали наибольшее влияние на предсказание метки класса (Аверченко А.Т.):

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|---|
| 1 | (9.74) | 1 | **А могли закусить и горяченьким: котлетками из рябчика, сосисочками в томате, грибочками в сметане… Да!! Слушайте а расстегаи?! Ах, Судаков, Судаков!.. Мне больше всего нравилось, что любой капитал давал тебе возможность войти в соответствующее место: есть у тебя 50 рублей пойди к Кюба, выпей рюмочку мартеля, проглоти десяток устриц, запей бутылочкой шабли, заешь котлеткой даньон, запей бутылочкой поммери, заешь гурьевской кашей, запей кофе с джинжером… Имеешь 10 целковых иди в «Вену» или в «Малый Ярославец». Обед из пяти блюд с цыпленком в меню целковый, лучшее шампанское 8 целковых, водка с закуской 2 целковых… А есть у тебя всего полтинник иди к Федорову или к Соловьеву: на полтинник и закусишь, и водки выпьешь, и пивцом зальешь… Эх, Федоров, Федоров!.. Кому это мешало?.. А летом в «Буфф» поедешь: музыка гремит, на сцене Тамара «Боккаччо» изображает… Помните? Как это она: «Так надо холить по-о-чку»… Ах, Зуппе! Ах, Оффенбах!.. Восточные человеки наговорились о своих делах, прислушиваются к разговору сенатора и директора завода. Слушают, слушают и полное непонимание на их лицах, украшенных солидными носами… На каком языке разговор?.. А «Маскотта»? «Сядем в почтовую карету, скорей»… А Джонсовская «Гейша»?.. «Глупо, наивно попала в сети я»… Ну!.. А «Луна-Парк»! А Айседора! А премьеры в Троицком или в Литейном!! А пуант с Фелисьеном и ужинами под румын, у воды!.. А аттракционы в Вилла Роде?.. А откровения психографолога Моргенштерна! Хе-хе… А разве лезло утром кофе в горло без «Петербургской Газеты»?! Да! С романом Брешки внизу! Как это он: «Виконт надел галифе, засунул в карман парабеллум, затянулся Боливаром, вскочил на гунтера, дал шенкеля и поскакал к авантюристу Петко Мирковичу!» Слова-то все какие подобраны, хе-хе… А «Сатирикон» по субботам! С утра торопишь Агафью чтобы сбегала за угол за журналом… А премьеры Андреевских пьес… Какое волнующее чувство. А когда художественники приезжали… И снова склоненные головы, и снова щемящий душу рефрен: Чем им мешало все это… Подходит билетер с книжечкой билетов и девица с огромным денежным ящиком.** | **7.38** | [CLS] а могли заку ##сить и горя ##чень ##ким : котле ##тками из ря ##б ##чика , соси ##со ##чками в тома ##те , гриб ##о ##чками в смета ##не … да ! ! слуша ##ите а рассте ##гаи ? ! ах , суда ##ков , суда ##ков ! . . мне больше всего нравилось , что любо ##и капитал давал тебе возможность вои ##ти в соответствующее место : есть у тебя 50 рубле ##и пои ##ди к кю ##ба , выпе ##и рюм ##о ##чку марте ##ля , прогло ##ти десяток устриц , запе ##и бутыло ##чко ##и ша ##бли , зае ##шь котле ##тко ##и дань ##он , запе ##и бутыло ##чко ##и пом ##мери , зае ##шь гур ##ьев ##ско ##и каше ##и , запе ##и кофе с джин ##жером … имеешь 10 цел ##ковых иди в « вену » или в « малы ##и ярослав ##ец » . обед из пяти блюд с цыплен ##ком в меню цел ##ковы ##и , лучшее шампанское 8 цел ##ковых , водка с заку ##ско ##и 2 цел ##ковых … а есть у тебя всего полтин ##ник иди к фе ##доров ##у или к солов ##ьеву : на полтин ##ник и заку ##сишь , и водки выпь ##ешь , и пив ##цом зал ##ье ##шь … эх , фе ##доров , фе ##доров ! . . кому это мешало ? . . а летом в « буф ##ф » поедешь : музыка гремит , на сцене там ##ара « бок ##кач ##чо » изображает … помните ? как это она : « так надо хол ##ить по - о - ч ##ку » … ах , зу ##ппе ! ах , офф ##ен ##бах ! . . восточные человек ##и наговорил ##ис ##ь о своих делах , прислуши ##ваются к разговору сенатора и директора завода . слушают , слушают и полное непонимание на их лицах , украшенных солидным ##и носами … на каком языке разговор ? . . а « маск ##от ##та » ? « сядем в почт ##ов ##ую карету , скор ##е ##и » … а джо ##нс ##ов ##ская « геи ##ша » ? . . « глупо , наивно попала в сети я » … ну ! . . а « луна - парк » ! а аи ##сед ##ора ! а премьеры в трои ##цком или в лит ##е ##ином ! ! а пу ##ант с фе ##лись ##еном и ужина ##ми под румын , у воды ! . . а аттракцион ##ы в вилла роде ? . . а откровения психо ##граф ##олог ##а морге ##н ##ште ##рна ! хе - хе … а разве лез ##ло утром кофе в горло без « петербург ##ско ##и газеты » ? ! да ! с романом бре ##шки внизу ! как это он : « виконт надел гали ##фе , засунул в карман пара ##белл ##ум , затянулся боли ##варом , вскочил на гун ##тера , дал шен |

# Этапы проекта и используемые технологии

В качестве основных этапов работы можно выделить следующие:

- сбор данных для обучения и размещение их на платформе Kaggle. Для работы с собранным датасетом будет использоваться Kaggle API.

- предобработка данных. Для приведения данных к формату, который удобен для обучения моделей, необходимо разделить тексты на отрывки, убрать лишние символы и т. п. Это выполняется с помощью регулярных выражений. Токенизация текстов будет производиться инструментами библиотеки transformers.

- обучение моделей производится во фреймворке Pytorch с использованием библиотеки transformers.

- логирование экспериментов и сохранение артефактов осуществляется с помощью сервиса WanDB.

- сохранение модели в формате ONNX для оптимизации инференса

- создание пайплайна обучения модели с помощью технологии DVC

- создание сервиса для инференса модели с помощью FastAPI

Кроме этого, будут использоваться следующие технологии:

GitHub — для контроля версий кода

Docker — для упаковки сервиса в контейнер

# Список литературы

1. Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang. How to Fine-Tune BERT for Text Classification? 2020
https://arxiv.org/pdf/1905.05583.pdf

2. Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. Fusing Label Embedding into BERT: An Efficient Improvement for Text Classification. 2021
https://aclanthology.org/2021.findings-acl.152.pdf

3. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, 2017
http://www.arxiv-vanity.com/papers/1703.01365/

4. Ganesh Jawahar, Benoıt Sagot, Djame Seddah. What does BERT learn about the structure of language?
https://aclanthology.org/P19-1356.pdf

5. Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li and Fei Wu. BertGCN: Transductive Text Classification by Combining GCN and BERT, 2022.
https://www.semanticscholar.org/reader/d22b109eb5089179f8bd48ef47513533890f6bf9

Приложение 1

| Писатель | Обучающая выборка | Тестовая выборка |
|---|---|---|
| Аверченко А.Т. | Том 1. Весёлые устрицы<br>Том 2. Круги по воде<br>Том 3. Чёрным по белому<br>Том 4. Сорные травы | Том 5. Чудеса в решете<br>Том 6. Отдых на крапиве |
| Акунин Б. | Чёрный город<br>Не прощаюсь<br>После тяжелой продолжительной болезни.<br>Время Николая II | Первая сверхдержава. История Российского государства. Александр Благословенный и Николай Незабвенный |
| Беляев А.Р. | Том 1. Остров Погибших Кораблей<br>Том 2. Последний человек из Атлантиды<br>Том 3. Человек-амфибия<br>Том 4. Властелин мира<br>Том 5. Прыжок в ничто<br>Том 8. Рассказы | Том 6. Звезда КЭЦ<br>Том 7. Человек, нашедший свое лицо |
| Булгаков М. | Черный маг (Черновики романа)<br>Записки юного врача<br>Белая гвардия<br>Морфий<br>Великий канцлер | Мастер и Маргарита |
| Бунин И.А. | Том 4. Повести и рассказы 1912-1916<br>Том 5. Рассказы 1917-1930'<br>Том 6. Жизнь Арсеньева<br>Том 7. Рассказы 1931-1952. Темные аллеи | Том 2. Рассказы 1892-1909<br>Том 3. Повести и рассказы 1909-1911<br>Том 9. Освобождение Толстого. О Чехове. Статьи |
| Гайдар А.П. | Том 1. Повести и рассказы<br>Том 3. Ранние и неоконченные произведения | Том 2. Повести, рассказы, фронтовые очерки |
| Гоголь Н.В. | Том 1. Вечера на хуторе близ Диканьки<br>Том 2. Миргород*<br>Том 3. Повести<br>Том 6. Мертвые души. Том 1<br>Том 7. Мертвые души. Том 2 | Том 4. Ревизор<br>Том 5. Женитьба. Драматические отрывки* |
| Гончаров В.А. | Том 1. Психо-машина<br>Том 6. Под солнцем тропиков. День Ромэна<br>Межпланетный путешественник<br>Том 3. Долина смерти | Том 4. Приключения доктора Скальпеля и фабзавука Николки<br>Том 5. Век гигантов |
| Горький М. | Случай с Евсейкой<br>Вор<br>Мои университеты<br>Гость<br>Жизнь Клима Самгина<br>Дед Архип и Лёнька<br>Исповедь<br>О евреях<br>Несвоевременные мысли: Заметки о революции и культуре | Фома Гордеев<br>Дело Артамоновых<br>Старуха Изергиль |
| Грин А.С. | Том 1. Рассказы 1906-1912<br>Том 4. Алые паруса. Романы<br>Том 5. Романы 1928-1930 | Том 2. Рассказы 1913-1916 |
| Довлатов С. | Том 1. Собрание прозы в четырех томах<br>Том 2. Собрание прозы в четырех томах<br>Том 4. Собрание прозы в четырех томах | Том 3. Собрание прозы в четырех томах |
| Достоевский Ф.М. | Том 1. Повести и рассказы 1846-1847<br>Том 2. Повести и рассказы 1848-1852 | Том 4. Произведения 1861-1866<br>Том 5. Преступление и наказание |

| | | |
|---|---|---|
| | Том 3. Село Степанчиково и его обитатели<br>Том 6. Идиот<br>Том 7. Бесы<br>Том 8. Вечный муж. Подросток<br>Том 11. Публицистика 1860-х годов<br>Братья Карамазовы | |
| Зощенко М.М. | Том 1. Разнотык<br>Том 2. Нервные люди<br>Том 3. Сентиментальные повести<br>Том 6. Шестая повесть Белкина<br>Том 7. Перед восходом солнца | Том 4. Личная жизнь<br>Том 5. Голубая книга |
| Ильф И., Петров Е. | Том 1. Двенадцать стульев<br>Том 4. Одноэтажная Америка<br>Том 5. Рассказы, очерки, фельетоны | Том 2. Золотой теленок<br>Том 3. Рассказы, фельетоны, статьи и речи |
| Казанцев А.П. | Том 1. Подводное солнце<br>Том 2. Сильнее времени<br>Том 3. Планета бурь. Фаэты<br>Том 4. Купол надежды<br>Том 7. Острие шпаги<br>Том 8. Мост дружбы<br>Том 9. Клокочущая пустота | Том 5. Льды возвращаются<br>Том 6. Пылающий остров |
| Катаев В.П. | Том 1. Рассказы и сказки<br>Том 2. Горох в стенку. Остров Эрендорф<br>Том 3. Растратчики. Время, вперед!<br>Том 4. Повести<br>Том 5. Белеет парус одинокий<br>Том 8. Почти дневник. Воспоминания | Том 6. Зимний ветер. Катакомбы<br>Том 7. Пьесы |
| Куприн А.И. | Том 1. Произведения 1889-1896<br>Том 2. Произведения 1896-1900<br>Том 3. Произведения 1901-1905<br>Том 6. Произведения 1914-1916 | Том 4. Произведения 1905-1907<br>Том 5. Произведения 1908-1913 |
| Лесков Н.С. | Том 1. Разбойник. Повести и рассказы<br>Том 2. Некуда<br>Том 3. Островитяне. Загадочный человек. Смех и горе<br>Том 4. Соборяне. Запечатленный ангел. Очарованный странник<br>Том 8. Пугало. Повести и рассказы<br>Том 9. Час воли божией. Повести и рассказы<br>Том 10. Воспоминания, статьи, очерки | Том 5. Захудалый род. Павлин. Детские годы<br>Том 6. Железная воля. Повести и рассказы<br>Том 7. Белый орел. Повести и рассказы |
| Лукьяненко С. | Пристань желтых кораблей (Сборник)<br>Планета, которой нет<br>Ночной Дозор<br>Дозоры: Последний Дозор. Новый Дозор. Шестой Дозор<br>Стеклянное море<br>Атомный сон (Сборник)<br>Принцесса стоит смерти | Книга гор: Рыцари сорока островов. Лорд с планеты Земля. Мальчик и тьма<br>Глубина: Лабиринт отражений. Фальшивые зеркала. Прозрачные витражи |
| Островский А.Н. | Том 1. Пьесы 1847-1854<br>Том 2. Пьесы 1856-1861<br>Том 3. Пьесы 1862-1864<br>Том 4. Пьесы 1865-1867<br>Том 5. Пьесы 1867-1870<br>Том 6. Пьесы 1871-1874<br>Том 9. Пьесы 1882-1885 | Том 7. Пьесы 1873-1876<br>Том 8. Пьесы 1877-1881 |
| Пастернак Б.Л. | Доктор Живаго | Апеллесова черта<br>Детство Люверс |

| | | |
|---|---|---|
| Паустовский К.Г. | Том 1. Романтики. Блистающие облака<br>Том 3. Повесть о лесах. Золотая роза<br>Том 4. Повесть о жизни. Книги 1-3<br>Том 5. Повесть о жизни. Книги 4-6<br>Том 7. Пьесы, рассказы, сказки 1941-1966 | Том 2. Черное море. Дым отечества<br>Том 6. Повести и рассказы 1922-1940 |
| Пелевин В. | Сочинения в двух томах. Том второй.<br>Поколение П<br>iPhuck-10<br>Искусство легких касаний | Сочинения в двух томах. Том первый<br>Синий фонарь<br>Empire V Бетмен Аполло |
| Пикуль В.С. | Реквием каравану PQ-17<br>Янычары<br>Богатство<br>Океанский патруль. Том 1. Аскольдовцы<br>Океанский патруль. Том 2. Ветер с океана | Крейсера. Ступай и не греши. Звезды над болотом<br>Париж на три часа |
| Пришвин М.М. | Том 1. В краю непуганых птиц. За волшебным колобком<br>Том 2. Кащеева цепь. Мирская чаша<br>Том 3. Журавлиная родина. Календарь природы<br>Том 4. Жень-шень. Серая Сова. Неодетая весна<br>Том 7. Натаска Ромки. Глаза земли | Том 5. Лесная капель. Кладовая солнца<br>Том 6. Осударева дорога. Корабельная чаща |
| Пушкин А.С. | Пиковая дама<br>Путешествие в Арзрум во время похода 1829 года<br>О народном воспитании<br>Метель<br>На углу маленькой площади<br>Станционный смотритель<br>Дубровский. Капитанская дочка (сборник)<br>История села Горюхино<br>История Пугачева | Арап Петра Великого<br>Марья Шонинг<br>Джон Теннер<br>Египетские ночи<br>Гробовщик<br>Барышня-крестьянка<br>Гости съезжались на дачу<br>Замечания о бунте |
| Салтыков-Щедрин М.Е. | Дневник провинциала в Петербурге<br>Мелочи жизни<br>История одного города. Господа Головлевы. Сказки | Губернские очерки<br>Господа ташкентцы |
| Серафимович А.С. | Том 1. Железный поток. Город в степи. Пески<br>Том 2. Рассказы, очерки, корреспонденции<br>Том 4. Скитания. На заводе. Очерки. Статьи | Том 3. В дыму орудий. В деревне. Дети |
| Сергеев-Ценский С.Н. | Том 1. Произведения 1902-1909<br>Том 2. Произведения 1909-1926<br>Том 3. Произведения 1927-1936<br>Том 4. Произведения 1941-1943<br>Том 10. Преображение России<br>Том 11. Преображение России<br>Том 12. Преображение России | Том 8. Преображение России<br>Том 9. Преображение России |
| Солженицин А.И. | Двести лет вместе. Часть первая<br>Двести лет вместе. Часть вторая<br>Архипелаг ГУЛАГ. 1918-1956: Опыт художественного исследования. Т. 1<br>Архипелаг ГУЛАГ. 1918-1956: Опыт художественного исследования. Т. 2<br>Архипелаг ГУЛАГ. 1918-1956: Опыт художественного исследования. Т. 3<br>Раковый корпус | В круге первом (т.1)<br>В круге первом (т.2)<br>Красное колесо. Узлы V - XX. На обрыве повествования |
| Стругацкие А. и Б. | Хромая судьба<br>Отель «У Погибшего Альпиниста»<br>Стажеры<br>Трудно быть богом | Улитка на склоне<br>За миллиард лет до конца света |

| | | |
|---|---|---|
| Толстой Л.Н. | Детство. Отрочество. Юность<br>Анна Каренина*<br>Война и мир. Книга 1<br>Война и мир. Книга 2 | Дьявол (сборник)<br>Воскресение. Повести. Рассказы* |
| Тургенев И.С. | Том 4. Повести и рассказы, статьи 1844-1854<br>Том 6. Дворянское гнездо. Накануне. Первая любовь<br>Том 7. Отцы и дети. Дым. Повести и рассказы 1861-1867<br>Том 8. Повести и рассказы 1868-1872<br>Том 9. Новь. Повести и рассказы 1874-1877<br>Том 10. Повести и рассказы 1881-1883 | Том 3. Записки охотника<br>Том 5. Рудин. Повести и рассказы 1853-1857 |
| Фадеев А.А. | Том 1. Разгром. Рассказы<br>Том 3. Молодая гвардия<br>Том 4. Очерки. Черная металлургия | Том 2. Последний из удэге |
| Фрай М. | Лабиринты Ехо. Том 1<br>Волонтеры вечности<br>Простые волшебные вещи<br>Сказки старого Вильнюса VII | Так [не] бывает<br>Зеленый. Том 3<br>Зеленый (темный). Том 3 |
| Фурманов Д.А. | Чапаев. Мятеж<br>Том 5. Путь к большевизму | Рассказы. Повести. Заметки о литературе |
| Чехов А.П. | Том 1. Рассказы, повести, юморески 1880-1882<br>Том 2. Рассказы, юморески 1883-1884<br>Том 3. Рассказы, юморески 1884-1885<br>Том 7. Рассказы, повести 1888-1891<br>Том 8. Рассказы, повести 1892-1894<br>Том 9. Рассказы, повести 1894-1897<br>Том 10. Рассказы, повести 1898-1903 | Том 5. Рассказы, юморески 1886<br>Том 6. Рассказы 1887 |
| Шукшин В.М. | Том 3. Рассказы 70-х годов | Том 2. Рассказы 60-х годов |

* Произведение было перенесено в другой набор данных — из обучающего в тестовый или наоборот

Приложение 2

| Пьеса | Проза |
|---|---|
| Аксюша. Раиса Павловна звали меня?<br>Карп. Так точное только теперь гости приехали, так они в саду.<br>Аксюша вынув из кармана письмо . Послушай, Карп Савельич, не можешь ли ты?..<br>Карп. Что вам угодно-с?<br>Аксюша. Передать. Ты уж знаешь кому.<br>Карп. Да как же, барышня? Теперь ведь уж словно как неловко. Правда ль, нет ли, у тетеньки такое есть желание, чтоб вам за барчонком быть.<br>Аксюша. Ну, не надо как хочешь. Отворачивается к окну.<br>Карп. Да уж пожалуйте. Для вас отчего же… Берет письмо.<br>Аксюша глядя в окно . Продала Раиса Павловна лес?<br>Карп. Продали Ивану Петрову. Все продаем-с, а чего ради?<br>Аксюша. Не хочет, чтоб наследникам осталось а деньги можно и чужим отдать.<br>Карп. Надо полагать-с. Мудрено сотворено.<br>Аксюша. Говорят, она эти деньги хочет за мной в приданое дать.<br>Карп. Дай-то бог!<br>Аксюша очень серьезно . Не дай бог, Карп Савельич!<br>Карп. Ну, как угодно-с. Я к тому, что все же лучше, пусть в приданое пойдут, чем туда же, куда и прочие.<br>Аксюша. Куда прочие… а куда же прочие?<br>Карп. Ну, это вам, барышня, и понимать-то невозможно, да и язык-то не поворотится сказать вам.<br>Алексей Сергеич идут. Отходит от двери.<br>Аксюша смотрит в окно, Буланов входит.<br>Явление второе<br>Аксюша, Буланов, Карп, потом Улита.<br>Буланов Карпу . Что ж, ты набил мне папиросы?<br>Карп. Никак нет-с.<br>Буланов. Отчего же нет? Ведь я тебе велел.<br>Карп. Мало что велели! А когда мне?<br>Буланов. Нет, уж вы здесь зазнались очень. Вот что. Я вот Раисе Павловне скажу.<br>Карп. Не скажете вы при них и курить-то боитесь.<br>Буланов. Боитесь… Чтоб были набиты! Не десять раз тебе говорить! Увидав Аксюшу, подходит к ней и очень развязно кладет ей на плечо руку.<br>Аксюша быстро обернувшись . Что вы! С ума сошли?<br>Буланов обидясь . Ах!! Извините! Что вы такой герцогиней смотрите, красавица вы моя?<br>Аксюша почти сквозь слезы . За что вы меня обижаете? Я вам ничего не сделала. Что я здесь за игрушка для всех? Я такой же человек, как и вы.<br>Буланов равнодушно . Нет, послушайте вы в самом деле мне нравитесь.<br>Аксюша. Ах, да мне-то что до этого за дело! | Сколько? вдруг заинтересовался Григорий Борисович.<br>Очень много. Думаю, больше ста.<br>И затем:<br>Вы под кроватью не смотрели?<br>Я посмотрю, сказал Григорий Борисович.<br>Писатель отодвинул кровать. Заглянул в кладовку. Порылся в ящиках стола.<br>Я завтра приду, сказал Ариэль.<br>С этого дня началась ежедневная пытка. Рано утром к нему заходил Ариэль:<br>Я только хотел спросить насчет машины.<br>Как сквозь землю провалилась, жаловался писатель.<br>Ничего, я вечером зайду.<br>В конце недели Григорий Борисович принял решение. Дневным автобусом поехал в Монтиселло. Зашел в игрушечный магазин Плейленд. Выбрал машину за сорок шесть долларов. Вернулся. Разыскал Ариэля и вручил ему большую, довольно тяжелую коробку.<br>Играй, сказал он.<br>Мальчик смутился.<br>Зачем? говорил он, срывая пластиковую ленту. Не беспокойтесь. Она найдется…<br>А потом:<br>К тому же это, в общем, другая машина. Капот не открывается.<br>Капот? переспросил Григорий Борисович. А я и не заметил. Колеса, думаю, на месте… Дверцы, руль…<br>Это не та машина, весело сказал Ариэль.<br>И положил ее в коробку. Поролоновые крепления вставил. Ленту приклеил на старое место.<br>Может, сойдет? упавшим голосом выговорил писатель.<br>Вы не беспокойтесь. Подумаешь, машина. У меня их штук двадцать пять. Правда, у той был капот. И фары.<br>У этой тоже фары.<br>У той были никелированные… Она найдется. Вы на кухне смотрели?<br>Смотрел.<br>А за плитой?<br>За плитой еще не смотрел.<br>Может, она там?<br>Григорий Борисович вынул из стола рейсшину. Долго водил ею за газовой плитой. Выкатил оттуда россыпь дряни, напоминавшей экскременты.<br>Не густо, сказал писатель.<br>Найдется, в который раз повторил Ариэль…<br>Короче, лето превратилось в ад. Ариэль появлялся, как тень отца в «Гамлете». Ужасом веяло на писателя от его слов:<br>Не беспокойтесь. Она найдется.<br>Писателю снились автомашины. Они съезжались к нему, беспомощному черные, громадные. Капоты их были угрожающе подняты. Никелированные фары сверкали.<br>Писатель обратился к Мишкевицеру. Тот сказал:<br>Да бросьте. Подумаешь, машина. У него их целый автопарк. |