ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК XXXXX

Отчет об исследовательском проекте на тему:
Определение авторства текстов

Выполнил:

студентка группы MMOBC21
Фофанова Татьяна Александровна

_____ _____
(подпись) (дата)

Принял руководитель проекта:

Аксенов Сергей

_____ _____
(подпись) (дата)

Москва 2023

Содержание

Аннотация

Задача определения авторства текстов является задачей классификации, признаковое пространство в которой совмещает как семантические и стилистические признаки текстов. Для проведения исследования был собран и выложен в открытый доступ набор данных с текстами русскоязычных писателей. Данная задача может быть решена как с помощью классических подходов машинного обучения, так и с использованием моделей глубинного обучения. В работе были рассмотрены оба подхода, наилучшие модели в каждом из них вошли в итоговый ансамбль моделей.

Ключевые слова

Классификация текстов, стилистическая близость текстов, семантическая близость текстов, метод интегрированных градиентов, NLP, BERT, SBERT, TF-IDF.

Введение

В данной работе 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.

Литературные произведения — это тексты большого объема, поэтому подход к решению задачи должен учитывать эту особенность.

Определение авторства текста тесно связано с определением стиля писателя. Стиль определяется как использованием характерной лексики, морфологии, синтаксиса, так и характерными сочетанием букв, который использует автор.

Основная задача, решаемая в работе — задача классификации текстовых данных; кроме основной, в работе решался ряд сопутствующих задач — кластеризация авторов по стилю, интерпретация полученной модели, агрегация предсказаний классификатора, написание сервиса на FastAPI для предсказания авторства текста.

Собранный набор данных представляет собой отрывки текстов на русском языке длиной не менее 2000 символов с соответствующей меткой автора. Отрывки из одного литературного произведения полностью входят либо в обучающий, либо в тестовый/валидационный набор данных, чтобы избежать завышения метрик качества. Объем полученной обучающей выборки — более 47 тыс. текстов, тестовой и валидационной — каждая более 9 тыс.

В качестве моделей, определяющих семантическую близость текстов, были взяты модели BERT и SBERT. Для определения стилистической близости — логистическая регрессия, построенная на символьных n-граммах и n-граммах частей речи (POS-теги).

Модели, использующие архитектуру трансформеров, были дообучены на собранном наборе данных, при этом веса инициализировались на основании модели ruBert-base, обученной на русскоязычных текстах из Википедии, новостях, части корпуса Taiga и книгах.

Подходы классического машинного обучения, использующие символьные и POS n-граммы и вектора TF-IDF, оказались более подходящими для решения данной задачи с точки зрения f1-меры качества моделей. Наибольшие значения метрики были достигнуты

при совмещении предсказаний двух моделей (BERT и логистическая регрессия) — f1-мера ансамбля моделей составила 0.845.

Для каждого текста тестового набора данных было вычислено векторное представление и с помощью метода t-SNE построено распределение объектов на 2-мерной плоскости, позволяющее визуально определить, насколько близки эти тексты для каждого автора, а также оценить, насколько авторы похожи между собой с точки зрения модели.

После обучения моделей был написан сервис, позволяющий определять авторство текстового отрывка с помощью различных моделей и определять степень уверенности в предсказании. Также сервис позволяет определять авторство большого текста, разбивая его на отрывки и агрегируя результат с помощью метода большинства.

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 языковую 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.

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.

Основная часть

Данные

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

Книга в EPUB-формате представляет собой набор файлов, каждый из которых содержит главу или отдельное короткое произведение (например, рассказ). Также в книге есть файлы стилей, обложка книги и другие вспомогательные файлы, которые не используются для обучения модели.

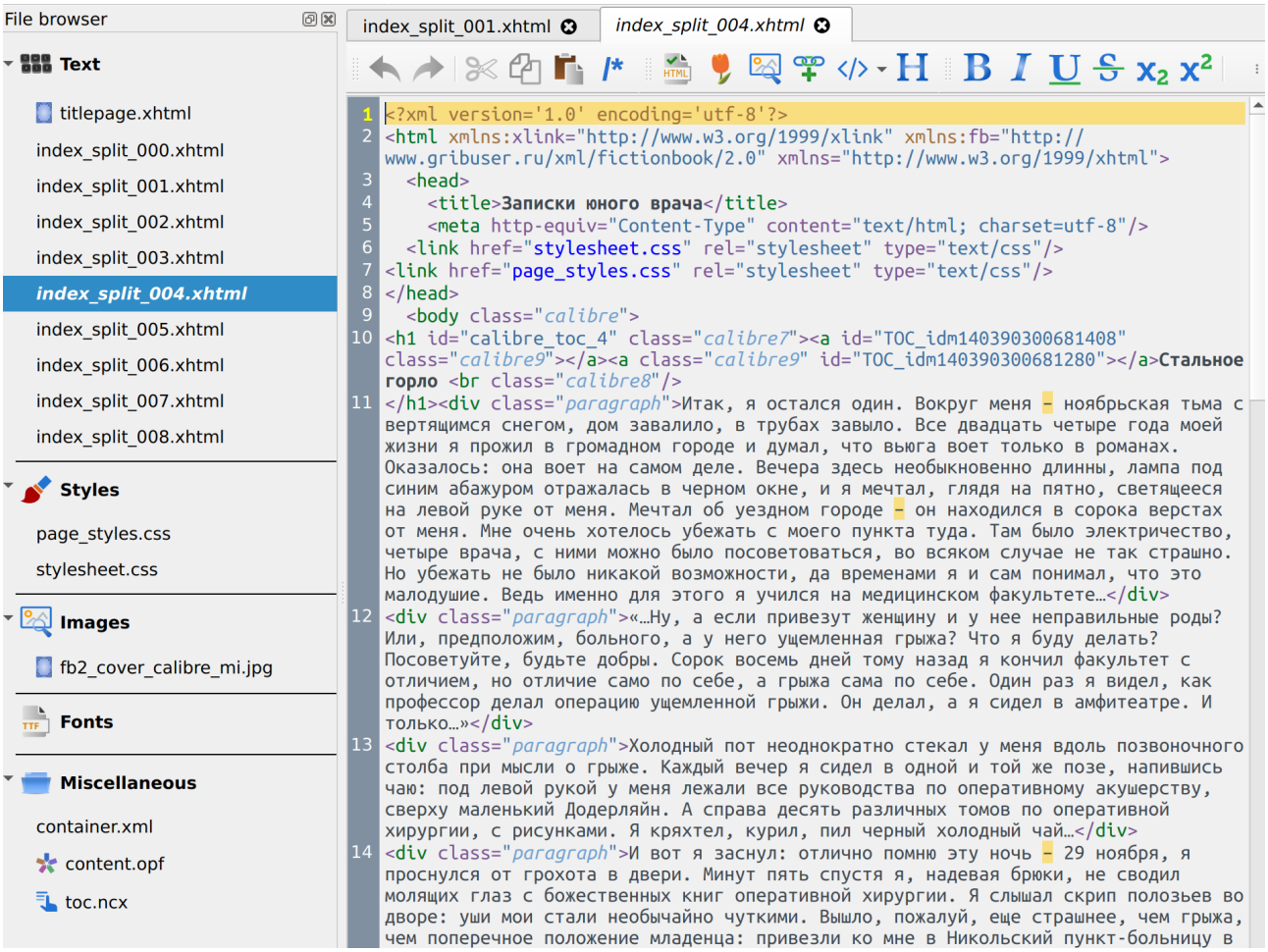

Рисунок 1. Книга в EPUB-формате

Для обучения модели текст главы нарезается на отрывки длиной не менее 2000 символов, состоящие из целых предложений. Так модель сможет обрабатывать произведение частями — для архитектуры трансформеров это важное ограничение.

Имея обученную модель, предсказывающую авторство для отрывков произведения, можно предсказать метку класса для всей главы / произведения — для этого необходимо агрегировать результаты, полученные для отдельных отрывков.

В обучающую выборку вошли 178 книг (томов, сборников и отдельных произведений), по ним были получены более 47 тыс. объектов (отрывков текстов); в тестовой выборке – 79 книг, что позволяет получить 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.

Далее тестовые данные были разделены на валидационные и тестовые, для того, чтобы можно было подобрать гиперпараметры моделей, не подгоняясь под тестовые данные. Пропорция валидационных и тестовых данных была выбрана 1:1.

Количество и объем произведений распределены по авторам неравномерно, наблюдается несбалансированность по классам. Это нужно учесть при подборе метрики качества классификации и иметь ввиду при анализе результатов классификации.

Увеличить выборку можно, вырезая отрывки текста с перекрытием. В работе планируется провести эксперимент с увеличенной обучающей выборкой.

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.

Будем считать, что главы, содержащие в тексте упоминание фамилии автора, являются шумом в данных, так как написаны другими писателями, и будем их игнорировать.

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

Главы, содержащие ссылки, короткие пояснения, сноски можно определить по объему текста — эти главы обычно короткие, меньше заданной минимальной длины отрывка.

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:

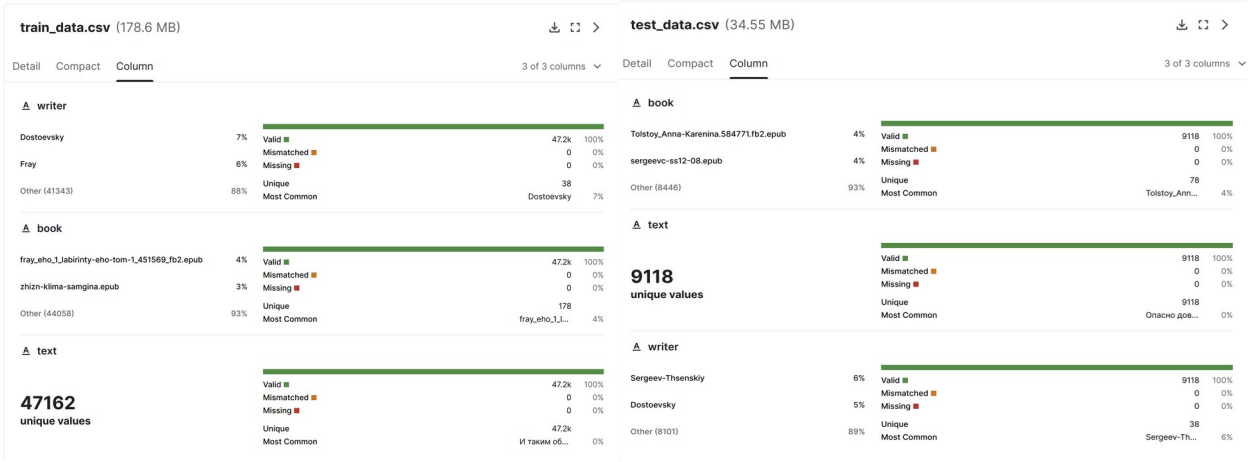

Рисунок 2. Описание набора данных на kaggle

Разведывательный анализ данных

Данные для обучения модели 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 сложно, удобнее 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.

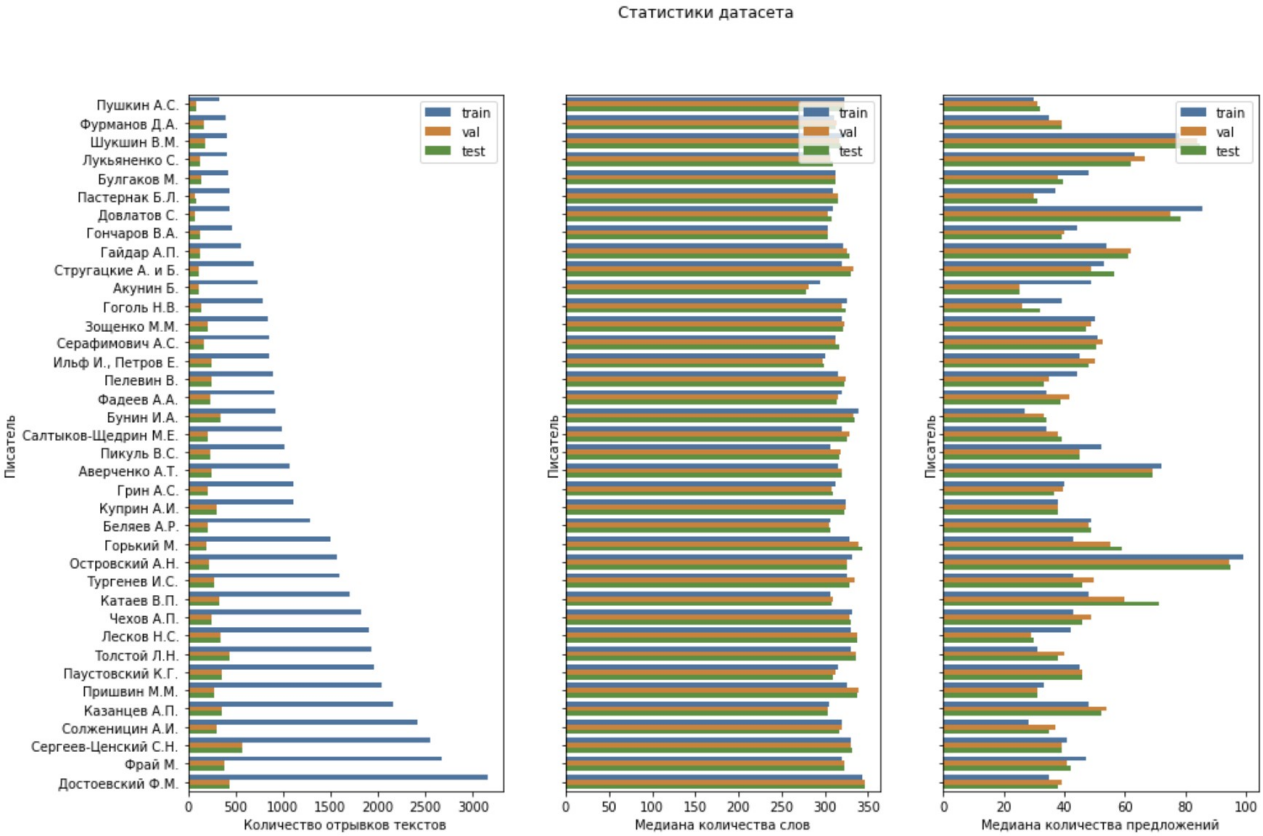

Рисунок 3. Характеристики обучающей, валидационной и тестовой выборок.

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

На рисунке 3 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 30 до 90 предложений).

Большое количество предложений характерно для пьес, где каждая реплика сопровождается указанием действующего лица. В качестве предобработки текстов предполагается провести эксперимент с удалением 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 рисунке 4, 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.

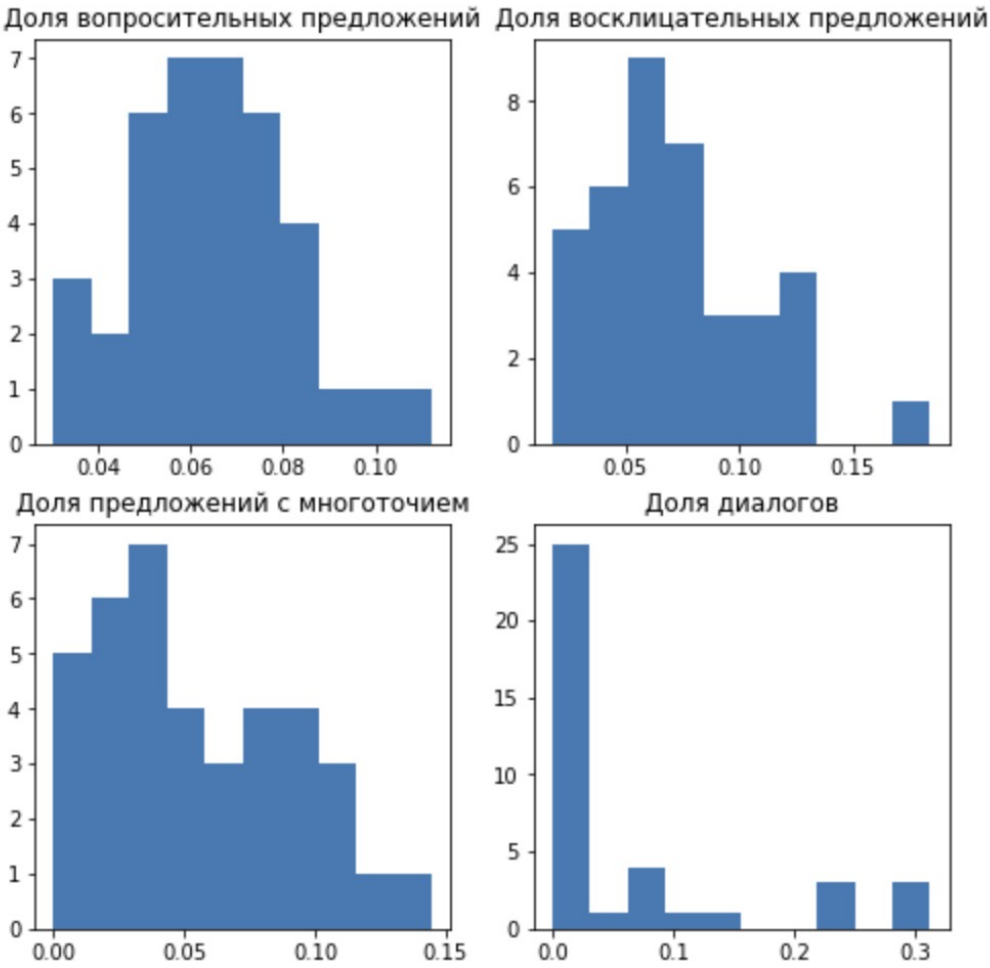

Рисунок 4. Распределение долей предложений разного типа в текстах авторов.

Построение базовой модели BERT

Задача классификации текстов — это классическая задача анализа текстовых данных, когда входной текстовой последовательности ставится в соответствие метка некоторого класса. В задаче с определением автора текста — метка писателя.

В качестве базовой модели классификации текстов была выбрана модель BERT (Bidirectional Encoder Representations from Transformers), обученная на большом корпусе русскоязычных текстов, и дообученная на примерах из собранного датасета.

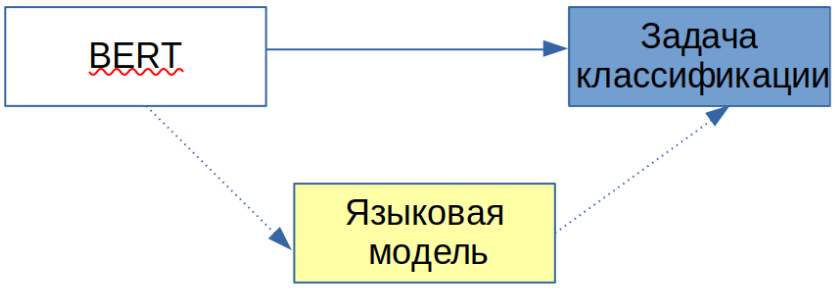

Рисунок 5. Использование предобученной модели BERT для классификации текстов.

Использование предобученных языковых моделей — стандартный подход для построения современных систем анализа текстовых данных, которая позволяет сократить использование вычислительных и временных ресурсов, используя уже имеющиеся у моделей знания о языке.

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 токенайзера 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 токенайзера, 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.

Наиболее вероятная метка класса вычисляется с помощью функции softmax, примененной к вектору Wh, где h - первый токен последнего слоя кодировщика BERT, соответствующему токену [CLS], несущему информацию о всей входной последовательности; W — параметры линейного слоя классификатора, преобразующего выход кодировщика в вектор, в котором каждая компонента соответствует метке класса.

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. При этом обучались веса всех слоев модели.

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 понять, каких писателей модель часто путает между собой:

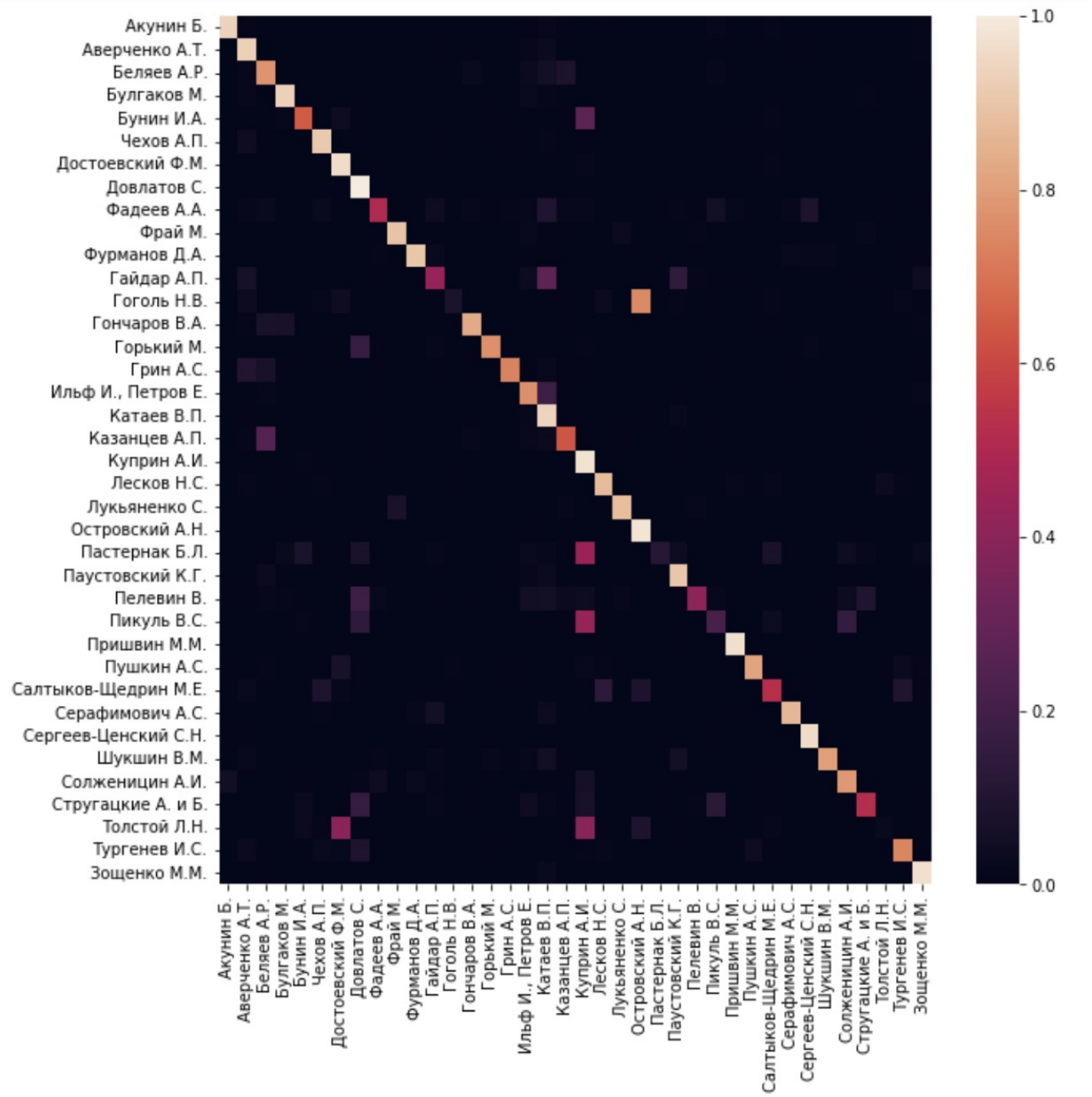

Рисунок 6. Матрица ошибок базовой модели.

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.

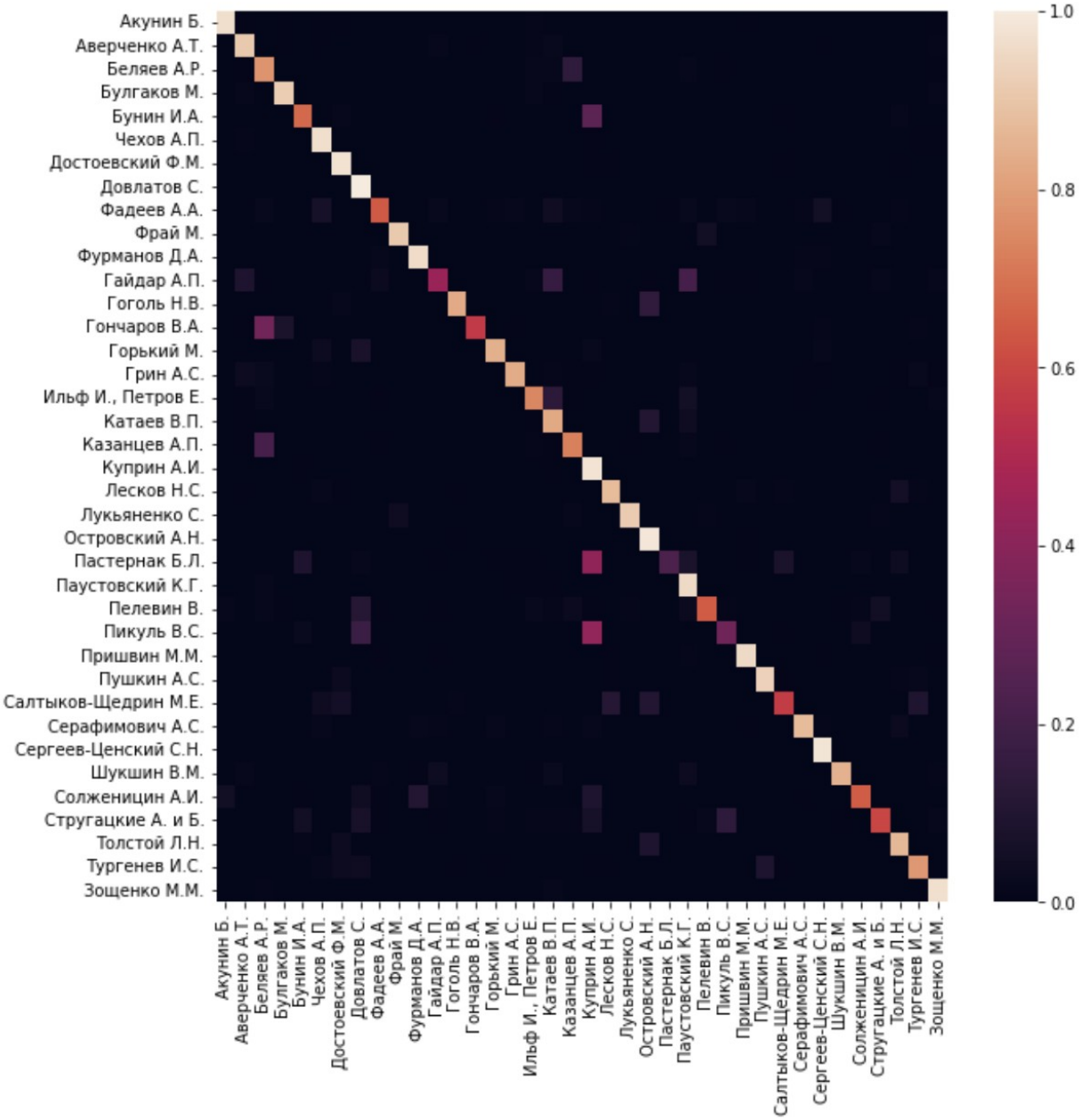

Рисунок 7. Матрица ошибок базовой модели после устранения недостатков разбиения.

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

1. Дообучение языковой модели на текстах из обучающей выборки

Для улучшения качества классификатора можно дообучить предобученную языковую модель на неразмеченных данных. На этом этапе языковая модель будет учиться на задаче предсказания маскированных слов.

После этого добавим к дообученной языковой модели классификатор и обучим их на имеющемся размеченном датасете.

Рисунок 8. Дообучение языковой модели BERT на домене перед обучением на задаче классификации.

В качестве данных, используемых для дообучения языковой модели, использовались отрывки из обучающей выборки. Доля маскированных слов — 15%.

В работе были проведены эксперименты с обучением языковой модели в течение 1, 2 эпох, с последующим обучением классификатора на основе BERT аналогично базовой модели.

Дообучение маскированной языковой модели на отрывках текстов писателей из решаемой задачи не привело к значительному улучшению качества классификации. Величина f1-score при макроусреднении в данном случае составила 0.796, при микроусреднении — 0.821.

Количество эпох обучения MLM на домене	F-score с макроусреднением	F-score с микроусреднением
0	0.793	0.821
1	0.796	0.821
2	0.789	0.820

2. 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 функции 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.

Рисунок 9. Представление объектов тестовой выборки с помощью метода t-SNE.

Классификатор на основе векторных представлений текстов, полученных с помощью SBERT, представляет собой линейный слой, получающий на вход эмбеддинг текста и выдающий логиты для каждого класса.

Данная модель не смогла улучшить качество классификации — значения метрик качества оказались ниже, чем у BERT: f1-score с макроусреднением составил 0.781, с микроусреднением — 0.794.

3. Логистическая регрессия

При построении классификатора на основе архитектуры BERT, использовался BPE тип токенизации исходного текста. При таком типе токенизации токенами являются слова или части слов, являющиеся наиболее частыми сочетаниями в корпусе обучающих данных.

Такой подход хорошо подходит для построения векторных представлений, учитывающих семантику. В случае определения авторства текстов, кроме семантики, а возможно и в большей степени, важно учитывать стилистические признаки текстов.

В частности, авторы могут использовать характерные для них буквенные сочетания. Для выявления таких сочетаний, разобьем тексты на 3, 4 и 5-буквенные n-граммы,

применим разные способы предобработки текстов - исключим из текста стопслова, пунктуацию символы, не относящиеся к кириллице.

Вычислим значения TF-IDF для всех n-грамм каждого текста и построим из низ векторные представления. На TF-IDF векторах обучим логистическую регрессию.

Предобработка текста	F-score с макроусреднением	F-score с микроусреднением
Только приведение к нижнему регистру и замена ё на е	0.821	0.839
Приведение к нижнему регистру, удаление стопслов, пунктуации	0.765	0.785
Приведение к нижнему регистру, удаление стопслов, пунктуации, символов, не относящихся к кириллице	0.751	0.770

Довольно неожиданно, что качество модели логистической регрессии без какой-либо предобработки текстов, оказалось выше качества классификатора, построенного на архитектуре BERT.

Кроме того, удаление стопслов, пунктуации и иностранных слов значительно снижает качество модели. Можно объяснить это следующим образом. Качество текстовых данных, получаемых моделью высокое, так как тексты взяты из книжного формата EPUB, где они прошли проверку на качество. Удаление пунктуации и стопслов лишает текст важной составляющей, по которой можно определить авторство текста. Поэтому для векторизации TF-IDF будем использовать тексты без предобработки, только приводя их к нижнему регистру.

У лучшей модели f1-score с макроусреднением достиг значения 0.821, с микроусреднением — 0.839.

Посмотрим на матрицу ошибок модели логистической регрессии:

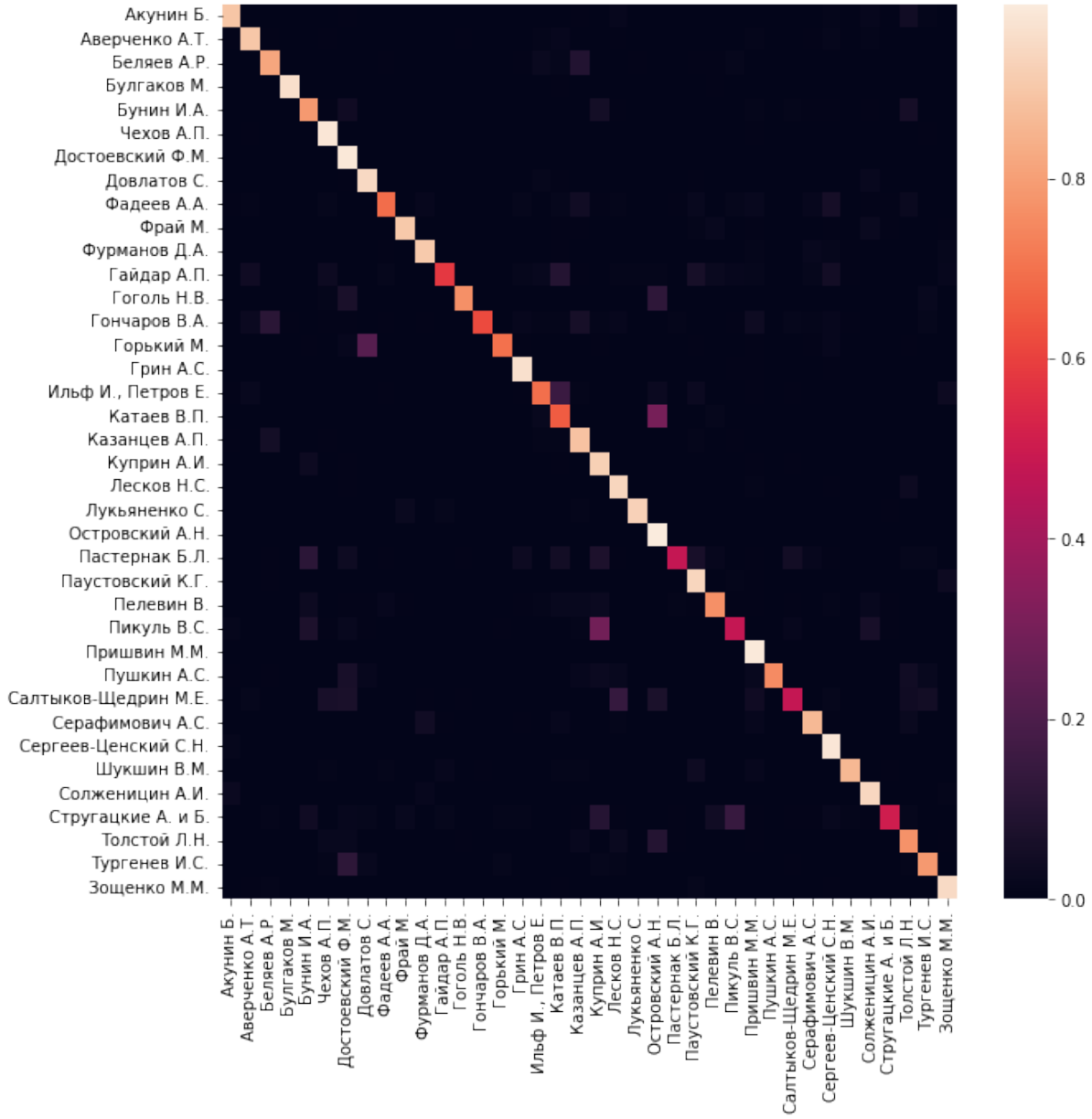

Рисунок 10. Матрица ошибок логистической регрессии, обученной на символьных n-граммах.

Логистическая регрессия, кажется, довольно хорошо справляется со всеми классами, хотя Паустовский, Пикуль, Стругацкие по-прежнему вызывают наибольшие трудности.

Но некоторые авторы, например Солженицин и Гончаров, оказались сложными в разной степени для логистической регрессии и BERT. Поэтому можно ожидать, что беггинг этих моделей позволит достичь лучшего качества, чем каждая из моделей по-отдельности.

Кроме символьных n-грамм, в качестве признаков можно использовать векторное представление последовательности POS-тегов (тегов, характеризующих части речи). В этом случае для каждого слова в тексте был поставлен в соответствие POS-тег и для этой последовательности проведена TF-IDF векторизация.

Полученный вектор был склеен в вектором, полученным на основании символьных n-грамм и на объединенном векторе была обучена логистическая регрессия.

Добавление POS-тегов в модель позволило увеличить качество логистической регрессии до f1-score с макроусреднением 0.825, с микроусреднением — 0.843.

Перебрав различные варианты разбиения на символьные и POS n-граммы, наилучшее качество оказалось у модели с диапазоном (3, 5) для символьных n-грамм, и (1, 3) для POS-тегов.

4. BERT с использованием нижних слоев энкодера

В базовом варианте классификатор, построенный на основе BERT, использует в качестве векторного представления текста первый токен с последнего слоя энкодера.

В качестве эксперимента можно строить векторное представление текста, учитывая не только последний слой, а взвешенную сумму выходов всех слоев энкодера.

Такой подход требует длительного времени на обучение моделей с различными весами слоев. В работе были проведены экперименты с построением векторных представлений на выходах:

- только первого слоя;

- первого и последнего слоя с весами 0.5;

- первого и последнего слоя с весами 0.05 и 0.95.

Все эти эксперименты оказались неудачными, f1-score с макроусреднением не превысил 0.500 и в дальнейшем этот подход использоваться не будет.

5. Ансамбль моделей

Поскольку объединение непохожих друг на друга моделей в ансамбль, как правило, позволяет улучшить общее качество классификации, то в окончательной модели будем использовать предсказания двух моделей с некоторыми весами.

Модели, входящие в ансамбль:

- логистическая регрессия, построенная на символьных n-граммах и учитывающая части речи;

- модель BERT.

Итоговое предсказание модели будет вычисляться следующим образом:

$$\hat{P}(y=C)=w*\hat{P}_{lr}(y=C)+(1-w)*\hat{P}_{BERT}(y=C)$$

Подбор веса w проводился по сетке значений от 0 до 1 с шагом 0.05.

Для подбора веса w будем использовать валидационную выборку, не совпадающую с тестовой, чтобы не подогнать результаты под тестовые данные.

Наилучшее качество было достигнуто при значении w, равного 0.75. При этом f1-score с макроусреднением составил 0.845, с микроусреднением — 0.860.

Результаты экспериментов

Результаты проведенных экспериментов приведены ниже:

Модель	F1 макро	F1 микро	Полнота	Точность
1. BERT	0.793	0.820	0.797	0.828
2. SBERT	0.781	0.794	0.787	0.816
3. Logreg (char)	0.821	0.839	0.813	0.856
4. Logreg (word)	0.756	0.782	0.743	0.816
5. Logreg (char+POS)	0.823	0.840	0.818	0.859
6. Ансамбль (1 + 5)	**0.845**	**0.860**	**0.839**	**0.878**

Наибольшее качество показал ансамбль моделей — логистической регрессии, построенной на символьных и POS n-граммах и BERT.

Визуализация

Построенная модель каждой входной последовательности ставит в соответствие метку класса, и, чтобы ответить на вопрос, почему модель сочла наиболее вероятной именно такую метку, нужно научиться интерпретировать предсказание модели.

Интерпретация предсказаний логистической регрессии

Логистическая регрессия представляет собой модель:

$$\hat{P}(y=C)=\frac{1}{1+e^{-\sum w_i x_i}}$$

Признаки, которым соответствуют большие положительные веса w_i, оказывают большое положительное влияние на вероятность $\hat{P}(y=C)$. И наоборот, большие отрицательные веса при признаках означают их негативное влияние на вероятность.

Признаки, наиважнейшие для каждого класса, как положительные, так и отрицательные, представлены в Приложении 3.

Среди положительных признаков часто встречаются имена главных героев произведений из обучающей выборки. Среди отрицательных признаков — сочетания частей речи с тексте, не характерное для автора. Также среди наиважнейших признаков часто встречаются сочетания знаков пунктуации, длинное тире, с которого начинаются диалоги.

Интерпретация предсказаний модели BERT

Поскольку коэффициенты в нейронных сетях, в отличие от линейных моделей, не имеют прямой интерпретации, то нужно использовать более сложные подходы.

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

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

Для оценки вклада каждого токена в предсказание модели, на ее вход подается две 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 совпадать с оригинальной, на m-м шаге — совпадать полностью.

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.

На каждом шаге интерполяции вычисляется изменение признака по сравнению с нулевым уровнем, и изменение выходного слоя модели. Если при изменении значения признака, предсказание модели меняется значительно, то он является важным для этой последовательности.

Каждый токен может вносить как положительный вклад в предсказание конкретного класса, так и отрицательный. Отрицательный вклад будет указывать на то, что такой токен не характерен для последовательностей этого класса.

Метод интегрированных 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

Рисунок 11. Анализ вкладов токенов в итоговое предсказание модели.

Заключение

В данной работе решалась задача определения авторства текстов с применением классических и глубинных методов машинного обучения, при этом совмещение моделей в ансамбль дало наилучшие результаты классификации. В качестве классических методов использовалось выделение из текста символьных n-грамм, n-грамм POS-тегов, построение на них TF-IDF векторов и обучение логистической регрессии. В качестве методов глубинного обучения были рассмотрены модели BERT-base, SBERT, с добавлением линейного слоя для решения задачи классификации. Исходные веса модели инициализировались на основании предобученной модели ruBert-base, размещенной на сервисе huggingface.

Для определения авторства текстов важно выделять из текстов как семантические, так и стилистические признаки. Признаки, выделяемые для обучения логистической регрессии, описывают стиль автора — характерные символьные сочетания, сочетания частей речи. Модель BERT в качестве признаков получает токены — слова/ части слов, они в большей части определяют семантическую составляющую текста.

Для решения задачи был собран и размещен в открытом доступе набор данных с текстами и метками авторов 38 русскоязычных писателей IIX — XXI веков.

Векторные представления, полученные с помощью модели были представлены графически на 2-мерной плоскости для определения близости текстов авторов относительно друг друга и оценки компактности векторных представлений текстов, относящихся к одному автору.

Модели, полученные в ходе экспериментов, были проанализированы с точки зрения влияния признаков на итоговое предсказание модели. Для линейной модели интерпретация не представляет сложности, но позволяет выделить наиболее важные для нее признаки, выделенные из текста.

Для нейронных сетей интерпретация результатов является гораздо более сложным вопросом. В данной работе влияние токенов в тексте на результат предсказания модели определялось с помощью метода интегрированных градиентов.

Качество моделей измерялось с помощью f1 метрики, рассчитанной как для каждого класса в отдельности, так с в среднем по всем классам с помощью макроусреднения.

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. Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019
https://arxiv.org/pdf/1908.10084.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 Том 2. Круги по воде Том 3. Чёрным по белому Том 4. Сорные травы	Том 5. Чудеса в решете Том 6. Отдых на крапиве
Акунин Б.	Чёрный город Не прощаюсь После тяжелой продолжительной болезни. Время Николая II	Первая сверхдержава. История Российского государства. Александр Благословенный и Николай Незабвенный
Беляев А.Р.	Том 1. Остров Погибших Кораблей Том 2. Последний человек из Атлантиды Том 3. Человек-амфибия Том 4. Властелин мира Том 5. Прыжок в ничто Том 8. Рассказы	Том 6. Звезда КЭЦ Том 7. Человек, нашедший свое лицо
Булгаков М.	Черный маг (Черновики романа) Записки юного врача Белая гвардия Морфий Великий канцлер	Мастер и Маргарита
Бунин И.А.	Том 4. Повести и рассказы 1912-1916 Том 5. Рассказы 1917-1930' Том 6. Жизнь Арсеньева Том 7. Рассказы 1931-1952. Темные аллеи	Том 2. Рассказы 1892-1909 Том 3. Повести и рассказы 1909-1911 Том 9. Освобождение Толстого. О Чехове. Статьи
Гайдар А.П.	Том 1. Повести и рассказы Том 3. Ранние и неоконченные произведения	Том 2. Повести, рассказы, фронтовые очерки
Гоголь Н.В.	Том 1. Вечера на хуторе близ Диканьки Том 2. Миргород* Том 3. Повести Том 6. Мертвые души. Том 1 Том 7. Мертвые души. Том 2	Том 4. Ревизор Том 5. Женитьба. Драматические отрывки*
Гончаров В.А.	Том 1. Психо-машина Том 6. Под солнцем тропиков. День Ромэна Межпланетный путешественник Том 3. Долина смерти	Том 4. Приключения доктора Скальпеля и фабзавука Николки Том 5. Век гигантов
Горький М.	Случай с Евсейкой Вор Мои университеты Гость Жизнь Клима Самгина Дед Архип и Лёнька Исповедь О евреях Несвоевременные мысли: Заметки о революции и культуре	Фома Гордеев Дело Артамоновых Старуха Изергиль
Грин А.С.	Том 1. Рассказы 1906-1912 Том 4. Алые паруса. Романы Том 5. Романы 1928-1930	Том 2. Рассказы 1913-1916
Довлатов С.	Том 1. Собрание прозы в четырех томах Том 2. Собрание прозы в четырех томах	Том 3. Собрание прозы в четырех томах

	Том 4. Собрание прозы в четырех томах	
Достоевский Ф.М.	Том 1. Повести и рассказы 1846-1847 Том 2. Повести и рассказы 1848-1852 Том 3. Село Степанчиково и его обитатели Том 6. Идиот Том 7. Бесы Том 8. Вечный муж. Подросток Том 11. Публицистика 1860-х годов Братья Карамазовы	Том 4. Произведения 1861-1866 Том 5. Преступление и наказание
Зощенко М.М.	Том 1. Разнотык Том 2. Нервные люди Том 3. Сентиментальные повести Том 6. Шестая повесть Белкина Том 7. Перед восходом солнца	Том 4. Личная жизнь Том 5. Голубая книга
Ильф И., Петров Е.	Том 1. Двенадцать стульев Том 4. Одноэтажная Америка Том 5. Рассказы, очерки, фельетоны	Том 2. Золотой теленок Том 3. Рассказы, фельетоны, статьи и речи
Казанцев А.П.	Том 1. Подводное солнце Том 2. Сильнее времени Том 3. Планета бурь. Фаэты Том 4. Купол надежды Том 7. Острие шпаги Том 8. Мост дружбы Том 9. Клокочущая пустота	Том 5. Льды возвращаются Том 6. Пылающий остров
Катаев В.П.	Том 1. Рассказы и сказки Том 2. Горох в стенку. Остров Эрендорф Том 3. Растратчики. Время, вперед! Том 4. Повести Том 5. Белеет парус одинокий Том 8. Почти дневник. Воспоминания	Том 6. Зимний ветер. Катакомбы Том 7. Пьесы
Куприн А.И.	Том 1. Произведения 1889-1896 Том 2. Произведения 1896-1900 Том 3. Произведения 1901-1905 Том 6. Произведения 1914-1916	Том 4. Произведения 1905-1907 Том 5. Произведения 1908-1913
Лесков Н.С.	Том 1. Разбойник. Повести и рассказы Том 2. Некуда Том 3. Островитяне. Загадочный человек. Смех и горе Том 4. Соборяне. Запечатленный ангел. Очарованный странник Том 8. Пугало. Повести и рассказы Том 9. Час воли божией. Повести и рассказы Том 10. Воспоминания, статьи, очерки	Том 5. Захудалый род. Павлин. Детские годы Том 6. Железная воля. Повести и рассказы Том 7. Белый орел. Повести и рассказы
Лукьяненко С.	Пристань желтых кораблей (Сборник) Планета, которой нет Ночной Дозор Дозоры: Последний Дозор. Новый Дозор. Шестой Дозор Стеклянное море Атомный сон (Сборник) Принцесса стоит смерти	Книга гор: Рыцари сорока островов. Лорд с планеты Земля. Мальчик и тьма Глубина: Лабиринт отражений. Фальшивые зеркала. Прозрачные витражи
Островский А.Н.	Том 1. Пьесы 1847-1854 Том 2. Пьесы 1856-1861 Том 3. Пьесы 1862-1864	Том 7. Пьесы 1873-1876 Том 8. Пьесы 1877-1881

	Том 4. Пьесы 1865-1867 Том 5. Пьесы 1867-1870 Том 6. Пьесы 1871-1874 Том 9. Пьесы 1882-1885	
Пастернак Б.Л.	Доктор Живаго	Апеллесова черта Детство Люверс
Паустовский К.Г.	Том 1. Романтики. Блистающие облака Том 3. Повесть о лесах. Золотая роза Том 4. Повесть о жизни. Книги 1-3 Том 5. Повесть о жизни. Книги 4-6 Том 7. Пьесы, рассказы, сказки 1941-1966	Том 2. Черное море. Дым отечества Том 6. Повести и рассказы 1922-1940
Пелевин В.	Сочинения в двух томах. Том второй. Поколение П iPhuck-10 Искусство легких касаний	Сочинения в двух томах. Том первый Синий фонарь Empire V Бетмен Аполло
Пикуль В.С.	Реквием каравану PQ-17 Янычары Богатство Океанский патруль. Том 1. Аскольдовцы Океанский патруль. Том 2. Ветер с океана	Крейсера. Ступай и не греши. Звезды над болотом Париж на три часа
Пришвин М.М.	Том 1. В краю непуганых птиц. За волшебным колобком Том 2. Кащеева цепь. Мирская чаша Том 3. Журавлиная родина. Календарь природы Том 4. Жень-шень. Серая Сова. Неодетая весна Том 7. Натаска Ромки. Глаза земли	Том 5. Лесная капель. Кладовая солнца Том 6. Осударева дорога. Корабельная чаща
Пушкин А.С.	Пиковая дама Путешествие в Арзрум во время похода 1829 года О народном воспитании Метель На углу маленькой площади Станционный смотритель Дубровский. Капитанская дочка (сборник) История села Горюхино История Пугачева	Арап Петра Великого Марья Шонинг Джон Теннер Египетские ночи Гробовщик Барышня-крестьянка Гости съезжались на дачу Замечания о бунте
Салтыков-Щедрин М.Е.	Дневник провинциала в Петербурге Мелочи жизни История одного города. Господа Головлевы. Сказки	Губернские очерки Господа ташкентцы
Серафимович А.С.	Том 1. Железный поток. Город в степи. Пески Том 2. Рассказы, очерки, корреспонденции Том 4. Скитания. На заводе. Очерки. Статьи	Том 3. В дыму орудий. В деревне. Дети
Сергеев-Ценский С.Н.	Том 1. Произведения 1902-1909 Том 2. Произведения 1909-1926 Том 3. Произведения 1927-1936 Том 4. Произведения 1941-1943 Том 10. Преображение России Том 11. Преображение России Том 12. Преображение России	Том 8. Преображение России Том 9. Преображение России
Солженицин А.И.	Двести лет вместе. Часть первая Двести лет вместе. Часть вторая Архипелаг ГУЛАГ. 1918-1956: Опыт художественного исследования. Т. 1 Архипелаг ГУЛАГ. 1918-1956: Опыт	В круге первом (т.1) В круге первом (т.2) Красное колесо. Узлы V - XX. На обрыве повествования

	художественного исследования. Т. 2 Архипелаг ГУЛАГ. 1918-1956: Опыт художественного исследования. Т. 3 Раковый корпус	
Стругацкие А. и Б.	Хромая судьба Отель «У Погибшего Альпиниста» Стажеры Трудно быть богом	Улитка на склоне За миллиард лет до конца света
Толстой Л.Н.	Детство. Отрочество. Юность Анна Каренина* Война и мир. Книга 1 Война и мир. Книга 2	Дьявол (сборник) Воскресение. Повести. Рассказы*
Тургенев И.С.	Том 4. Повести и рассказы, статьи 1844-1854 Том 6. Дворянское гнездо. Накануне. Первая любовь Том 7. Отцы и дети. Дым. Повести и рассказы 1861-1867 Том 8. Повести и рассказы 1868-1872 Том 9. Новь. Повести и рассказы 1874-1877 Том 10. Повести и рассказы 1881-1883	Том 3. Записки охотника Том 5. Рудин. Повести и рассказы 1853-1857
Фадеев А.А.	Том 1. Разгром. Рассказы Том 3. Молодая гвардия Том 4. Очерки. Черная металлургия	Том 2. Последний из удэге
Фрай М.	Лабиринты Ехо. Том 1 Волонтеры вечности Простые волшебные вещи Сказки старого Вильнюса VII	Так [не] бывает Зеленый. Том 3 Зеленый (темный). Том 3
Фурманов Д.А.	Чапаев. Мятеж Том 5. Путь к большевизму	Рассказы. Повести. Заметки о литературе
Чехов А.П.	Том 1. Рассказы, повести, юморески 1880-1882 Том 2. Рассказы, юморески 1883-1884 Том 3. Рассказы, юморески 1884-1885 Том 7. Рассказы, повести 1888-1891 Том 8. Рассказы, повести 1892-1894 Том 9. Рассказы, повести 1894-1897 Том 10. Рассказы, повести 1898-1903	Том 5. Рассказы, юморески 1886 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? Карп. Так точное только теперь гости приехали, так они в саду. Аксюша вынув из кармана письмо . Послушай, Карп Савельич, не можешь ли ты?.. Карп. Что вам угодно-с? Аксюша. Передать. Ты уж знаешь кому. 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. Аксюша. Ну, не надо как хочешь. Отворачивается к окну. Карп. Да уж пожалуйте. Для вас отчего же… Берет письмо. Аксюша глядя в окно . Продала Раиса Павловна лес? Карп. Продали Ивану Петрову. Все продаем-с, а чего ради? Аксюша. Не хочет, чтоб наследникам осталось а деньги можно и чужим отдать. Карп. Надо полагать-с. Мудрено сотворено. Аксюша. Говорят, она эти деньги хочет за мной в приданое дать. Карп. Дай-то бог! Аксюша очень серьезно . Не дай бог, Карп Савельич! Карп. Ну, как угодно-с. Я к тому, что все же лучше, пусть в приданое пойдут, чем туда же, куда и прочие. Аксюша. Куда прочие… а куда же прочие? Карп. Ну, это вам, барышня, и понимать-то невозможно, да и язык-то не поворотится сказать вам. Алексей Сергеич идут. Отходит от двери. Аксюша смотрит в окно, Буланов входит. Явление второе Аксюша, Буланов, Карп, потом Улита. Буланов Карпу . Что ж, ты набил мне папиросы? Карп. Никак нет-с. Буланов. Отчего же нет? Ведь я тебе велел. Карп. Мало что велели! А когда мне? Буланов. Нет, уж вы здесь зазнались очень. Вот что. Я вот Раисе Павловне скажу. Карп. Не скажете вы при них и курить-то боитесь. Буланов. Боитесь… Чтоб были набиты! Не десять раз тебе говорить! Увидав Аксюшу, подходит к ней и очень развязно кладет ей на плечо руку. Аксюша быстро обернувшись . Что вы! С ума сошли? Буланов обидясь . Ах!! Извините! Что вы такой герцогиней смотрите, красавица вы моя? 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. Буланов равнодушно . Нет, послушайте вы в самом деле мне нравитесь. Аксюша. Ах, да мне-то что до этого за дело!	Сколько? вдруг заинтересовался Григорий Борисович. Очень много. Думаю, больше ста. И затем: Вы под кроватью не смотрели? Я посмотрю, сказал Григорий Борисович. Писатель отодвинул кровать. Заглянул в кладовку. Порылся в ящиках стола. Я завтра приду, сказал Ариэль. С этого дня началась ежедневная пытка. Рано утром к нему заходил Ариэль: Я только хотел спросить насчет машины. Как сквозь землю провалилась, жаловался писатель. Ничего, я вечером зайду. 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. Играй, сказал он. Мальчик смутился. Зачем? говорил он, срывая пластиковую ленту. Не беспокойтесь. Она найдется… А потом: К тому же это, в общем, другая машина. Капот не открывается. Капот? переспросил Григорий Борисович. А я и не заметил. Колеса, думаю, на месте… Дверцы, руль… Это не та машина, весело сказал Ариэль. И положил ее в коробку. Поролоновые крепления вставил. Ленту приклеил на старое место. Может, сойдет? упавшим голосом выговорил писатель. Вы не беспокойтесь. Подумаешь, машина. У меня их штук двадцать пять. Правда, у той был капот. И фары. У этой тоже фары. У той были никелированные… Она найдется. Вы на кухне смотрели? Смотрел. А за плитой? За плитой еще не смотрел. Может, она там? Григорий Борисович вынул из стола рейсшину. Долго водил ею за газовой плитой. Выкатил оттуда россыпь дряни, напоминавшей экскременты. Не густо, сказал писатель. Найдется, в который раз повторил Ариэль… Короче, лето превратилось в ад. Ариэль появлялся, как тень отца в «Гамлете». Ужасом веяло на писателя от его слов: Не беспокойтесь. Она найдется. Писателю снились автомашины. Они съезжались к нему, беспомощному черные, громадные. Капоты их были угрожающе подняты. Никелированные фары сверкали. Писатель обратился к Мишкевицеру. Тот сказал: Да бросьте. Подумаешь, машина. У него их целый автопарк.

Приложение 3

Автор	Позитивные	Негативные
Акунин Б.	'фандо', 'ндори', 'дорин', 'омано', 'андор', 'дори', 'фанд', 'орин', 'ндор', 'эраст'	'pl', 'num ciph nonlex', '\n— ', 'nonlex nonlex nonlex', 'nonlex num ciph', 'part pro', ' и ', ' — ', 'conj', 'pro'
Аверченко А.Т.	'!! ', '?! ', 'anum ciph', 'nonlex anum ciph', ' я. ', ' я.', 'nonlex part nonlex', 'л я', 'ысако', 'крыса'	'nonlex num ciph', ' не ', 'pr nonlex', 'не ', ' он', ' и ', ' не', 'adv', '\n— ', ' — '
Беляев А.Р.	'вагне', 'вагн', 'агнер', 'гнер', ' ганс', 'ганс', ' слон', 'аркер', 'ркер', 'прест'	' чт', 'ciph', 'num ciph', 'то ', 'adv pro conj', 'part nonlex', '\n — ', 'pr', 'adv pro', ' — '
Булгаков М.	'пилат', 'илат', ' пила', 'оланд', 'её ', 'волан', ' её', 'иколк', 'урато', 'олан'	'pro pro', 'все', 'nonlex', 'pro pl', 'nonlex nonlex', 'adv', '\n— ', ' на', 'pro', ' — '
Бунин И.А.	' — ', ':\n—', ':\n— ', ', —', ', — ', ' — и', '\n— ', ' — и ', '— и', '— и '	'самг', 'амгин', 'амги', 'мгин', '.. ', 'num', '...', 'nonlex conj', 'pro', 'nonlex'
Чехов А.П.	' и ', ' по', ', то ', 'pr conj', '... п', 'о... ', ' егор', '! с', 'горуш', '! ска'	'pl', 'но ', 'pro pro', ', — ', ', —', '.. ', 'nonlex', 'adv', '\n— ', ' — '
Достоевский Ф.М.	' — ', ', —', ', — ', '\n— ', 'чтоб ', 'тоб ', '.\n—', '.\n— ', 'наро', ' фом'	' ска', 'сказа', 'азал', 'казал', 'pr pr', 'ой ', ' сказ', '...', 'nonlex conj', 'nonlex'
Довлатов С.	'...', '...\n', '..\n', '\n— ', ' — ', ':\n— ', ':\n—', 'nonlex pro', '... ', ', — '	'амг', 'самги', 'самг', 'мгин', 'амгин', 'амги', 'pro', 'pr', 'part', 'conj'
Фадеев А.А.	'мечик', 'олег', ' олег', 'нерет', 'мечи', 'ечик', ' мечи', 'езнев', 'знев', 'орозк	'nonlex anum ciph', ' я ', ', — ', ', —', 'part', 'adv', 'nonlex anum', '\n— ', 'nonlex', ' — '
Фрай М.	'сэр', ' сэр', ' сэ', ' коф', 'коф', 'сэр ', 'эр ', ' сэр ', 'ффи', 'джу'	'он ', 'adv conj', 'nonlex nonlex', 'pr', 'conj nonlex', ' он', 'pro', ' и ', 'nonlex', 'conj'
Фурманов Д.А.	'чапа', 'чапае', 'апаев', ' чапа', 'паев', 'чап', ' чап', 'апае', ' што', ' што '	'conj pro', 'nonlex conj pro', 'conj', 'nonlex pro', 'nonlex conj', 'pro', 'pro nonlex', ' — ', 'nonlex', 'pr'
Гайдар А.П.	' лбов', 'лбов', 'лбо', ' яшк', 'натка', ' лбо', 'яшк', 'имк', 'натк', ' натк'	'adv pro part', ' мо', 'pro pro', ' в ', '\n— ', 'nonlex nonlex', 'pr', 'pro', ' — ', 'nonlex'
Гоголь Н.В.	'чичик', 'чичи', '!» ', '.\n«', 'ичико', 'вши ', '»\n«', 'ичик', ' коза', 'коза']	' а ', ' я ', ' эт', ', — ', ', —', 'nonlex conj pro', 'conj pro', '\n — ', 'nonlex', ' — '
Гончаров В.А.	'тьк', 'петьк', 'етьк', 'тька', '.. ', 'етька', '!..', ' петь', '?..', 'петь'	'\n— ', 'nonlex num ciph', 'nonlex nonlex', 'pro nonlex', 'то ', 'pr nonlex', 'nonlex', 'nonlex conj', 'conj', ' — '

Горький М.	' - ', 'клим', ' — ', 'амги', 'мгин', 'амгин', 'самг', 'самги', 'амг', 'мги'	'pl pl nonlex', 'num ciph', 'nonlex', 'conj', 'adv pro', 'ciph', 'pl pl', 'pr', 'part', 'pro']
Грин А.С.	'моськ', ' гент', 'геник', 'мось', 'оськ', ' я ', ', с', ' мось', ' ге', 'conj nonlex'	', — ', ', —', ' на', 'pro', 'nonlex nonlex nonlex', 'adv', '\n— ', 'part', ' — ', 'pr
Ильф И., Петров Е.	'остап', 'стап', 'тап', 'адамс', '». ', 'мы ', 'вееви', 'дамс', 'амс', ' мы '	'part conj', '...', 'part adv', 'pr', 'pr pro', 'conj', 'ciph', '\n— ', 'pro', ' — '
Катаев В.П.	'петя', 'етя', ' петя', 'етя ', 'петя ', '. он', 'аврик', 'врик', '. о', ' ван'	'!.. ', 'pro nonlex', 'conj', ', — ', ', —', 'ся ', '.. ', '!..', '\n— ', ' — '
Казанцев А.П	'сиран', 'ирано', 'сира', 'вилен', 'виле', ' сира', 'ферма', 'иран', ' виле', ' фаэ'	'adv conj adv', ' эт', 'nonlex pr', ', и', ', и ', 'nonlex nonlex nonlex', 'adv', 'conj', ' — ', ' и '
Куприн А.И.	' — ', ':\n— ', ':\n—', '\n— ', ' олес', ', точ', 'pl nonlex pl', 'очно ', 'лихон', 'ихони'	'pr', '...\n', 'pro nonlex', ' всё ', 'всё ', 'сё ', ' всё', 'всё', 'nonlex', '...'
Лесков Н.С.	'ою ', ' эт', 'ною', 'кою', 'настя', 'вечал', 'айнер', 'твеча', 'отвеч', 'твеч'	'. и', '. и ', ' всё', 'всё', ' он', ', — ', ', —', '\n— ', 'nonlex', ' — '
Лукьяненко С.	'тири', 'вамп', 'вампи', 'ампир', 'мпир', ' я.', ' вамп', 'ампи', '\nя ', ' тири'	'num ciph', '\n— ', 'ciph', 'nonlex nonlex nonlex', 'pr', 'pr pro', 'conj', ' — ', 'nonlex nonlex', 'pro'
Островский А.Н.	'в. ', 'ов. ', 'на. ', 'ов.', 'на.', 'ина. ', 'nonlex part', '\nне', 'ов\n', 'ина.'	'нул', 'nonlex pro nonlex', 'pr', ' по', 'nonlex nonlex nonlex', 'ал ', 'nonlex conj', '\n— ', 'adv', ' — '
Пастернак Б.Л.	' юр', ' лар', 'лар', 'рееви', 'юри', ' лара', ' юри', 'рий а', 'лара', ' не'	'nonlex conj', 'ciph nonlex', 'num', 'conj', 'num ciph', ' — ', 'part', 'ciph', 'nonlex', 'adv'
Паустовский К.Г.	'.\nя ', 'бату', '.\nя', 'nonlex adv pro', '\nя ', ' бату', 'лес', 'одесс', 'пушки', 'десс'	' ко', 'adv', 'кот', 'ciph', 'отор', ' кот', 'котор', 'кото', ' кото', ' — ']
Пелевин В.	'атарс', 'саш', 'тарс', 'тарск', 'саша', 'татар', ' саш', 'тата', 'саша ', 'сэм'	' и ', 'nonlex pro', 'nonlex pr', 'nonlex pl', 'num ciph', 'nonlex pl nonlex', 'pr num ciph', 'nonlex pr nonlex', 'pl nonlex', 'nonlex'
Пикуль В.С.	'...', '..\n', 'ломин', '... ', 'оломи', '.. ', 'nonlex conj', ' фин', 'ломи', 'солом'	'pro conj', 'num', 'pro', ' бы', 'nonlex nonlex nonlex', '\n— ', 'pr conj', ' и ', ' — ', 'conj'
Пришвин М.М.	'лпато', ', и ', 'патов', ', и', 'пато', ' бобр', 'алпат', 'лпат', 'бобр', 'алпа'	'adv', '!..', ' всё', ' кото', 'всё', '.. ', ' ва', 'nonlex', ' не', ' — '
Пушкин А.С.	'угаче', 'пугач', 'угач', 'гачев', 'гаче', 'pro pr', 'бровс', 'злоде', 'злод', ' злод'	'то ', 'nonlex part', 'part nonlex', ' это', 'это', 'nonlex conj', ' эт', 'adv', 'part', 'nonlex'
Салтыков-Щедрин М.Е.	'рокоп', 'ежели', 'ежел', 'проко', 'лупов', 'еже', 'окоп',	'nonlex pro', ' всё ', 'всё ', 'сё ', 'вич', 'num ciph nonlex', '

	'жели ', ' ежел', 'жели'	всё', 'всё', 'nonlex', 'pro'
Серафимович А.С.	'!..', ' гля', 'рабоч', 'pro part', 'рабо', 'и, ', ' гл', 'або', '!.. ', ' раб'	'conj nonlex pr', 'казал', 'parenth nonlex', 'num ciph', 'азал', 'nonlex', 'pr', '\n— ', 'pro', ' — '
Сергеев-Ценский С.Н.	'.. ', '!.. ', '!..', '?.. ', '?..', ' леня', 'ивенц', 'венце', 'ливен', 'бабае'	'. н', ' ра', ' всё', 'всё', 'а. ', 'nonlex pro', 'и. ', 'nonlex', '\n — ', ' — '
Шукшин В.М.	'adv pro nonlex', 'nonlex part', '-то', '-то?', 'nonlex part nonlex', 'одовн', ' тоже', ' тож', 'нязев', 'тоже'	'nonlex', 'adv pro part', 'pro', 'pro conj', 'pl', 'nonlex nonlex', '\n— ', ' — ', 'pr', 'conj'
Солженицин А.И.	' — ', ' . ', ' евре', 'евре', ' — и', ' — и ', ' евр', '— и', '— и ', ' лаг'	'что', ' ска', ' сказ', '...\n', 'pr pro', ' ск', ' я ', '...', 'ciph nonlex num', 'nonlex'
Стругацкие А. и Б.	'румат', 'рума', 'умата', ' рума', ' юр', 'мата', 'рум', 'вик', 'юра', 'викто'	'comp', 'pr', 'pro', 'num ciph', 'nonlex conj nonlex', 'nonlex adv', 'conj nonlex', 'ciph', 'pr pro', 'nonlex'
Толстой Л.Н.	' граф', 'пьер', ' пьер', 'хлюдо', 'нехлю', 'нехл', 'ехлюд', 'хлюд', 'ехлю', 'людов'	'pro pr', 'adv nonlex', 'nonlex nonlex pro', ' всё ', 'всё ', 'сё ', ' всё', 'adv', 'всё', 'nonlex'
Тургенев И.С.	'санин', 'еич', ' не', 'омолв', 'олвил', 'лвил', ' да ', 'молв', 'твино', 'итвин'	'.\n— ', 'pr', 'nonlex adv pro', 'adv pro', ', — ', ', —', 'nonlex conj', '\n— ', 'nonlex', ' — '
Зощенко М.М.	'. и ', '. и', '.\ни', '.\ни ', '\na ', '.\na ', '.\нн', 'и.\n', ', гов', '\ни	'nonlex', ', а', 'pro part', ', а ', 'pr pr', '\n— ', 'nonlex nonlex', ', и', ', и ', ' — '