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

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 004.8

Отчет об исследовательском проекте на тему:

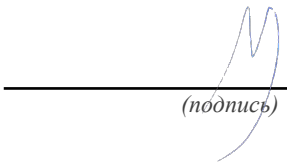
“Разработка метода оценки сходства датасетов”

Выполнил:
студент группы БПМИ-205
Копылов Олег Иванович

(подпись) 6 мая 2023 *(дата)*

Принял руководитель проекта:

Валентин Андреевич Малых
кандидат технических наук, инженер ключевых проектов
ООО "Техкомпания Хуавэй"



(подпись) 6 мая 2023 *(дата)*

Москва 2023

Аннотация

В настоящее время существует большое число доступных открытых датасетов. Но еще большее число датасетов закрытых, которые существуют в рамках отдельных компаний. Не всегда легко понять, какие из существующих языковых моделей подойдут для конкретного закрытого датасета. Например, в случае, если обучаться на закрытом датасете невозможно или нецелесообразно.

В рамках курсовой работы будет разработан метод, который помогает справиться с этой проблемой. Это метод оценки сходства датасетов. При условии сходства датасетов A и B, модель, эффективно работающая на датасете A, также эффективно работает и на датасете B.

Метод оценки сходства датасетов представляет собой функцию, которая принимает два датасета и возвращает численное значение, коррелирующее с изменением качества при переходе с одного датасета на другой. Таким образом, если открытый датасет A схож с закрытым датасетом B, и проверена эффективность модели на открытом датасете A, то эта модель эффективно работает и на закрытом датасете B.

Датасеты предполагаются текстовыми, рассматривается задача бинарной классификации. Проведено две серии экспериментов, первая серия относится к определению положительности или негативности отзыва, вторая серия – к классификации содержимого сообщений или электронных писем как спам и не спам. В каждой серии экспериментов рассмотрено два датасета и 10 языковых моделей.

Ключевые слова: NLP, языковые модели, сходство датасетов, закрытые датасеты, перенос качества.

Оглавление

Определения

Качество – в рамках этой работы под качеством подразумевается F1 score или F1 score и accuracy.

Прямое тестирование – тестирование, при котором модель, обученная на обучающей выборке некоторого датасета, тестируется на тестовой выборке этого же датасета.

Перекрестное тестирование – тестирование, при котором модель, обученная на обучающей выборке одного датасета, тестируется на тестовой выборке другого датасета.

Сходство датасетов – свойство двух датасетов A и B, при котором модель, эффективно работающая на датасете A, также эффективно работает и на датасете B. В рамках данной работы сходство датасетов понимается двумя способами: сходство датасетов в первом смысле и во втором смысле.

Сходство датасетов в первом смысле – свойство двух датасетов A и B, при котором модель, обученная на датасете A и показывающая хорошие результаты при тестировании на A, также показывает хорошие результаты при тестировании и на B. В этом случае модель обучается только на датасете A, тогда как на датасете B происходит лишь тестирование.

Сходство датасетов во втором смысле – свойство двух датасетов A и B, которое состоит в следующем: **если** модель M после обучения на датасете A показывает хорошие результаты при тестировании на A, **то** эта же самая модель M, но уже обученная на датасете B, покажет хорошее качество при тестировании на датасете B.

NLP (Natural Language Processing, обработка естественного языка) – это направление искусственного интеллекта, связанное с распознаванием, обработкой и генерацией человеческой речи.

Трансфо́рмер (Transformer) – архитектура глубоких нейронных сетей, представленная в 2017 году исследователями из Google Brain.

Языковая модель – это распределение вероятностей по последовательностям слов.

Постановка задачи

Рассмотрим тривиальную, на первый взгляд, задачу: дан текстовый датасет, нужно подобрать наиболее подходящую для него модель из N возможных языковых моделей. Очевидное решение – сначала обучить на данном датасете все N моделей, а затем выбрать ту, которая показывает наилучшее качество при тестировании. Однако такое решение, к сожалению, применимо *не всегда*.

Во-первых, данный датасет может быть очень маленьким. Из-за этого обучаться непосредственно на нём не целесообразно: ввиду слишком маленькой обучающей выборки модель не запомнит закономерность.

Во-вторых, датасет может не обладать всей информацией, требуемой для тестирования или обучения на нём. Например, датасет состоит только из текстов отзывов, каждый из которых нам нужно классифицировать либо как положительный, либо как отрицательный. В самом датасете нет информации о том, какой отзыв является положительным, а какой – отрицательным, даны только тексты отзывов. В таком случае обучаться на датасете не представляется возможным.

Итак, в обоих случаях нельзя обучить модель непосредственно на данном датасете. Следовательно, необходимо обучить её на каком-то другом датасете. Наверное, его выбор имеет смысл производить не случайно, а на основании некоторого метода оценки сходства датасетов. Напомним, что при наличии сходства между двумя датасетами модель, хорошо работающая на одном из двух датасетов, также эффективна и на оставшемся датасете.

Благодаря методу оценки сходства датасетов станет возможным подобрать для датасета (из условия задачи выше) другой более удобный и в то же время схожий датасет. Затем обучить N языковых моделей на новом подобранном датасете и после этого выбрать модель с наилучшим качеством, таким образом решив поставленную задачу.

Однако совершенно не понятно, по каким принципам определять сходство датасетов, на что опираться. Подобные исследования не проводились ранее. Поэтому мне бы хотелось провести такое исследование.

Цель исследования – разработать метод оценки сходства текстовых датасетов или обосновать, почему такой метод разработать невозможно. Разработанный метод оценки сходства датасетов должен быть практически полезным, то есть помогать подобрать подходящие модели для датасета.

Для достижения вышеуказанной цели поставлены следующие **задачи**:

1) изучение научных работ о языковых моделях,
2) поиск моделей и датасетов для проведения экспериментов,
3) определение методики проведения экспериментов,
4) проведение экспериментов,
5) визуализация результатов экспериментов,
6) формулирование гипотез о том, на что должны опираться методы оценки сходства датасетов исходя из результатов пунктов 4-5,
7) проверка сформулированных гипотез с помощью разработки соответствующих методов оценки сходства датасетов,

8) тестирование эффективности разработанных методов,
9) визуализация и оценка полученных результатов,
10) анализ проделанной работы и формулирование итогового вывода,
11) формулирование направлений дальнейших исследований.

Актуальность и значимость

В настоящее время NLP стремительно развивается, и в важности этой области искусственного интеллекта нет сомнения: повсеместно используются поисковые системы, распознавание речи, опеределение спама и многое другое.

При решении задач NLP, как я объяснил в разделе “постановка задачи”, иногда не представляется возможным подобрать подходящую языковую модель для данного текстового датасета посредством обучения на данном датасете. В этом случае требуется использовать некоторый метод оценки сходства датасетов. Таким образом, результат моей работы может быть использован в задачах обработки естественного языка для решения описанной выше проблемы.

Научная **новизна** заключается в том, что будет проведено исследование по поиску и разработке метода оценки сходства датасетов, такое исследование не проводилось ранее.

Теоретическая часть 1 – Обзор источников

Рассматриваемые датасеты содержат тексты на естественном языке, поэтому необходимо использовать языковые модели. Я изучил несколько статей и использовал описанные в этих статьях языковые модели при проведении своих экспериментов.

В статье [1] 2018 года представлена языковая модель BERT (Bidirectional Encoder Representations from Transformers). По сравнению с известными ранее моделями, преимущество BERT в том, что его можно дообучить для решения рассматриваемой задачи (например, классификации отзывов, выделения спама или ответа на вопросы) при этом не внося существенных изменений в архитектуру модели.

По мнению авторов, “Существенное ограничение стандартных языковых моделей – это их однонаправленность: она ограничивает выбор архитектур, которые могут быть использованы во время предобучения”. Данное ограничение может препятствовать успешному решению тех задач, в которых важно учитывать контекст слева и справа от рассматриваемого токена (слова или части слова). Поэтому BERT является двунаправленной моделью.

BERT работает по следующему принципу: текст разбивается на токены, от которых берутся эмбеддинги (представления токенов в виде числовых векторов), после чего эмбеддинги постепенно обновляются с помощью механизма self-attention для контекста. Обучение BERT состоит из двух этапов: предобучение (pre-training) и дообучение (fine-tuning).

Во время предобучения модель учится решать две задачи. Первая – это “masked language modeling” (MLM): небольшая часть токенов заменяется на токен [MASK], и модель учится угадывать, какие токены стоят на местах токена [MASK]. Вторая – “next sentence prediction” (NSP): модель угадывает, следует ли после предложения ***A*** предложение ***B***. Для предобучения используются два корпуса слов: English Wikipedia и BooksCorpus. Таким образом, в результате предобучения модель получает некоторое общее знание языка, которое не зависит от конкретной решаемой задачи.

А уже во время fine-tuning модель обучается конкретной задаче обработки естественного языка на соответствующих этой задаче датасетах. Например, если нужно обучить модель классифицировать отзывы как “позитивные” или “негативные”, то на этапе fine-tuning будет использоваться датасет с отзывами, для каждого из которых указано, является он позитивным или негативным.

При тестировании на GLUE (The General Language Understanding Evaluation benchmark) [2] модель BERT показала GLUE score, равный 80,5%.

Это было абсолютным улучшением на 7,7% по сравнению с существовавшими на тот момент передовыми языковыми моделями.

Как я узнал из статьи [2], GLUE – это набор инструментов для оценки производительности моделей на различных задачах NLU (natural language understanding). NLU является частью NLP.

GLUE фокусируется на таких задачах, как ответы на вопросы, анализ тональности текста и *textual entailment* (определение то, следует ли из одного утверждения другое). Авторы GLUE не создавали датасеты сами с нуля, а использовали уже существующие наборы данных, которые, как написано в статье, “считаются в NLP сообществе сложными и интересными”.

Также BERT тестировали на датасете SQuAD (The Stanford Question Answering Dataset) [3], где проверяется способность модели ответить на вопрос словами из текста. Абсолютное увеличение F1 score составило 1,5% на SQuAD v1.1 и 5,1% на SQuAD v2.0.

Помимо BERT есть и другие модели, которые я использовал в своих экспериментах. Например, RoBERTa: A Robustly Optimized BERT Pretraining Approach. Она описана в статье [4]. Авторы утверждают, что BERT был значительно недообучен на этапе предобучения и предлагают новую методику предобучения, которая вносит следующие изменения:

1) более длительное предобучение модели, с большим batch size и на большем количестве данных;
2) удаление задачи “next sentence prediction”;
3) обучение на более длинных последовательностях;
4) динамическое изменение шаблона маскировки (он применяется в задаче “masked language modeling”).

В результате применения этих изменений удалось добиться улучшения качества на GLUE, SQuAD и RACE.

В RACE (The ReAding Comprehension from Examinations) [5] модель отвечает на вопросы по тексту, к каждому вопросу предлагается четыре вариант ответа и нужно выбрать один верный.

Еще одна модель – ELECTRA из статьи [6] показывает лучшие по сравнению с BERT результаты за счет видоизменения задачи “masked language modeling”. В отличие от BERT, где некоторые токены заменялись на [MASK] и нужно было предсказать стоящий за [MASK] токен, теперь некоторые токены заменяются на правдоподобные альтернативы. Затем модель должна предсказать, был ли токен заменен на правдоподобную альтернативу.

В статье “Language-agnostic BERT Sentence Embedding” [7] представлена модель LaBSE, которая имеет схожую с BERT архитектуру и поддерживает 109

языков. Модель демонстрирует отличные результаты при выполнении перевода предложений на английский и выполнении задач интеллектуального анализа текста. Предобученная модель представлена в открытом доступе на TensorFlow Hub [8]. Там же можно найти и другие модели. Их я использую далее в работе.

Теоретическая часть 2 – Понятие сходства датасетов

Поскольку сходство датасетов мы определяем в терминах совпадения хорошо работающих моделей, то, чтобы понять что-то про сходство двух рассматриваемых датасетов, следует сначала взять N моделей и затем проверить, верно ли, что модели, хорошо работающие для датасета A, хорошо работают и для датасета B. При этом “совпадение хорошо работающих моделей” можно понимать двумя способами.

Первый способ – это про *перенос качества* с одного датасета на другой. То есть модель, обученная на датасете A и показывающая хорошие результаты при тестировании на A, также показывает хорошие результаты при тестировании и на B. Отмечу, что в этом случае модель обучается только на датасете A, тогда как на датасете B происходит лишь тестирование.

Второй способ – это о том, что обученная на датасете A модель показывает хорошее качество на датасете A и эта же самая модель, но уже обученная на датасете B, показывает хорошее качество при тестировании на датасете B. В отличие от первого способа понимания, здесь уже нет как такового *переноса* качества: перед тестированием на каком-то из датасетов модель обучается на нём же.

Теоретическая часть 3 – Метод оценки сходства датасетов

Каждый метод оценки сходства датасетов представляет собой функцию, которая принимает два датасета и возвращает численное значение. Возвращаемое численное значение тем больше, чем лучше происходит переход с первого датасета на второй.

Метод оценки сходства двух датасетов должен коррелировать с изменением качества при переходе с одного датасета на другой. Под изменением качества подразумевается относительное или абсолютное изменение f1_score.

Практическая часть 1 – Используемые модели

По описанным в теоретической части соображениям, для проведения исследования нужно несколько моделей, а не одну. Количество рассматриваемых моделей, то есть N, было выбрано равным 10. Используются следующие модели:

1) BERT
2) ROBERTA
3) ALBERT
4) ELECTRA
5) DistilBERT
6) MobileBERT
7) LaBSE
8) talkheads_ggelu_bert (BERT with Talking-Heads Attention and Gated GELU)
9) LAMBERT
10) tn_bert (A compressed BERT model using tensor networks)

Эти модели я загружаю с TensorFlow Hub [8], они уже предобучены. Их можно настроить под конкретную задачу. Я решил остановиться на задаче *бинарной классификации*, так как она является наиболее простой из задач классификации.

Модели были специально подобраны таким образом, чтобы они были похожи на BERT, принадлежали одному с ним семейству NLP моделей. Благодаря этому мы ожидаем, что модели будут вести себя схожим образом, то есть, например, изменение качества при переходе с одного датасета на другой будет примерно одинаковым для большинства моделей.

Практическая часть 2 – Выбор датасетов для проведения серий экспериментов

Серией экспериментов будем называть обучение и тестирование моделей на двух датасетах A и B.

Первая серия экспериментов посвящена бинарной классификации отзывов на позитивные и негативные. Рассматриваемых датасетов два: первый с отзывами об отелях, второй с отзывами о фильмах. С моей точки зрения, эти датасеты могут быть схожими, потому что связаны с одной и той же задачей: разделением отзывов на положительные и отрицательные.

Для второй серии экспериментов я взял два датасета для решения уже другой задачи бинарной классификации, а именно определения спама. В первом

датасете содержатся SMS, а во втором – e-mail письма. Для всех них указано, является ли SMS или e-mail спамом.

Отмечу, что в обеих сериях экспериментов датасеты специально подобраны таким образом, чтобы можно было исследовать сходство этих датасетов в первом смысле, то есть чтобы перекрестное тестирование имело смысл.

Таким образом, сходство датасетов в первом смысле возможно только между датасетами из одной и той же серии экспериментов (потому что только в этом случае перекрестное тестирование имеет смысл). Для сходства датасетов во втором смысле такое ограничение отсутствует. Сходство датасетов во втором смысле возможно как между датасетами из *одной* серии экспериментов, так и между датасетами из *разных* серий экспериментов.

Практическая часть 3 – Обучение и тестирование моделей

Ноутбуки с обучением и тестированием моделей представлены в гитхаб-репозитории. Для каждой модели создана отдельная папка (название папки соответствует названию модели), где представлены ноутбуки с обучением модели (для всех серий экспериментов), тестированием (в том числе и перекрестным). В каждой папке присутствует свой readme файл, в котором описана навигация по файлам внутри папки.

Эксперименты я решил проводить в Colab, где Google предоставляет свои вычислительные ресурсы, так как мне хотелось бы не перегружать свой компьютер вычислениями. При использовании Colab можно либо загружать файлы с компьютера и использовать их на протяжении сессии (в новой сессии файлы придется загружать заново), либо загрузить файлы в Google Drive один раз и с помощью специальной библиотеки получить доступ к файлам из Google Drive. Я выбрал второй способ, так как загружать файлы для каждой сессии в Colab – это дополнительные расходы Интернет-трафика и времени.

Обучение проводилось так, чтобы оно занимало примерно одинаковое количество времени для каждой модели и каждого датасета. Исходя из этого, при обучении моделей допускалось различное число эпох.

Рассмотрим на примере модели BERT **основные этапы** экспериментов, затрагивающих обучение и тестирование с участием этой модели.

Для первой серии экспериментов, где были взяты датасет с отзывами об отелях и датасет с отзывами о фильмах:

1) **предварительная подготовка данных**: датасет загружается из Google Drive, представляется в виде DataFrame со столбцами “DATA_COLUMN”, “LABEL_COLUMN”. В первом из этих столбцов содержится текст

отзыва, а во втором – флаг, который равен 1, если отзыв положительный (позитивный), и 0, если отзыв отрицательный (негативный).

2) **разделение отзывов на положительные и отрицательные**: формируем два датафрейма, в одном только положительный отзывы, а в другом – только отрицательные.
3) **создание обучающей и тестовой выборки**: с помощью датафреймов из предыдущего пункта формируются датафрейм с обучающей выборкой и датафрейм с тестовой выборкой. В рамках **первой** серии экспериментов обучающая и тестовая выборки сделаны сбалансированными, то есть в каждой выборке число позитивных отзывов примерно равно числу негативных отзывов. Однако, сразу оговоримся, во второй серии экспериментов уже будет присутствовать дисбаланс классов в обучающей и тестовой выборках, мы обсудим способ решения этой проблемы далее в работе.
4) **создание языковой модели**: с TensorFlow Hub [8] скачиваем препроцессор и кодировщик для BERT. Создаем модель из трех слоев: первый слой – это входной слой, второй слой отвечает за препроцессинг и кодировку, третий слой отвечает за дропаут (нужен для предотвращения переобучения, установили dropout rate равным 0.1), применение полносвязного слоя и сигмоидальной функции активации.
5) **обучение языковой модели**: в METRICS указываем, значения каких метрик хотелось бы отслеживать в процессе обучения модели. Выбраны метрики accuracy, precision, recall.

   В model.compile в качестве optimizer выбран Adam, в качестве loss – бинарная кросс-энтропия, в качестве метрик передаются метрики из METRICS (см. выше).

   С помощью model.fit происходит дообучение модели на рассматриваемом датасете (в рамках первой серии экспериментов это либо датасет с отзывами об отелях, либо датасет с отзывами о фильмах), где модель учится определять позитивность или негативность отзыва.
6) **тестирование модели на том же датасете (прямое тестирование)**:

   тестирование на тестовой выборке того же датасета, на котором модель обучали, происходит в том же ноутбуке, где происходило обучение.
7) **сохранение обученной модели:** обученная модель первоначально сохраняется в Google Drive. Данный этап необходим для последующего перекрестного тестирования, в ходе которого модель, обученная на одном датасете, тестируется не на нём же, а на другом датасете.

   Проблема в том, что у бесплатного Google Drive место в хранилище ограничено 15 Гб, чего не достаточно для сохранения всех моделей.

Поэтому дополнительно пришлось использовать привязанный к университетскому аккаунту Yandex Disk, размер хранилища в котором составляет 1 Тб.

Успешно завершив перекрестное тестирование в Colab, я скачиваю из Google Drive модель локально на компьютер, после чего загружаю на Yandex Disk (чтобы стало возможным предоставить к ней доступ по ссылке). Теперь можно удалить модель из Google Drive.

8) **тестирование в рамках серии экспериментов, в т. ч. перекрестное**:

В отдельный ноутбук вынесено всё тестирование в рамках одной серии экспериментов: как прямое тестирование (когда обучение модели и её тестирование происходит на одном и том же датасете), так и перекрестное тестирование (когда обучение модели происходит на одном датасете, а тестирование – на другом). Например, в ходе перекрестного тестирования в рамках первой серии экспериментов модель, обученная определять позитивность или негативность отзыва об отеле, тестируется на датасете с отзывами о фильмах. Перекрестное тестирование также подразумевает “переход” и в обратную сторону: модель, обученная определять позитивность или негативность отзыва о фильме, тестируется на датасете с отзывами об отелях. Датасеты здесь берутся из одной и той же серии экспериментов, потому что только в этом случае имеет смысл проводить перекрестное тестирование – см. замечание ниже.

Из Google Drive загружаются две уже дообученные модели (одна модель BERT дообучена на датасете с отзывами об отелях, другая же модель BERT дообучена на датасете с отзывами о фильмах).

Результаты тестирования сохраняются в Google Drive в виде csv файла.

Замечание:

Напомним, что перекрестное тестирование имеет отношение к первому способу понимания сходства датасетов. Можно заметить, что перекрестное тестирование имеет смысл проводить не для всех пар датасетов. Например, оно не целесообразно, если датасет A состоит из сообщений с маркером, является ли сообщение спамом, тогда как датасет B представляет собой отзывы с указанием, является ли отзыв положительным или отрицательным. Дело в том, что модель, обученная на датасете A, научится распознавать спам. При тестировании этой модели на датасете B она скажет, какие отзывы по её мнению спам, а какие – нет. Не понятно, как оценивать такой результат: для отзывов дана только информация о позитивности или негативности отзыва и нет информации о том, является ли отзыв спамом. Для решения этой проблемы в рамках одной серии экспериментов датасеты специально подобраны так, чтобы перекрестное тестирование имело смысл.

Для второй серии экспериментов, которая посвящена классификации сообщений и электронных писем на спам и не спам, можно выделить те же самые этапы с некоторыми корректировками:

1) **предварительная подготовка данных**: аналогично первой серии экспериментов, за исключением того, что вместо “LABEL_COLUMN” вторая колонна называется “IS_SPAM”. Записанное в ней значение равно 1, если сообщение является спамом, и 0 иначе.
2) **разделение сообщений или электронных писем на спам и не спам**: аналогично первой серии экспериментов.
3) **создание обучающей и тестовой выборки**: в отличие от первой серии экспериментов, здесь уже будет присутствовать дисбаланс классов: сообщений со спамом будет заметно меньше, чем сообщений, которые не являются спамом. Это вынужденная мера, так как в используемых датасетах, если поддерживать баланс классов в тестовой и обучающей выборках, обучающая выборка получалась слишком маленькой. Далее в работе я учитываю имеющийся дисбаланс классов, подбираю соответствующие метрики качества и тд, поэтому проблемы в этом нет.
4) **создание языковой модели**: аналогично первой серии экспериментов.
5) **обучение языковой модели**: аналогично первой серии экспериментов.
6) **тестирование модели на том же датасете (прямое тестирование)**: аналогично первой серии экспериментов.
7) **сохранение обученной модели**: аналогично первой серии экспериментов.
8) **тестирование в рамках серии экспериментов, в т. ч. перекрестное**: аналогично первой серии экспериментов, за исключением примера перекрестного тестирования: модель, обученная определять, является ли текст SMS спамом, тестируется на определение спама в электронных письмах.

Для остальных 9 моделей, в целом, эксперименты были проведены схожим образом: также проводятся две серии экспериментов по аналогичному алгоритму.

Практическая часть 4 – Гитхаб-репозиторий

В гитхаб репозитории (на данный момент приватном) выложены все файлы с кодом.

Описание папок вида "название папки – описание её содержимого":

1. datasets – используемые в работе датасеты
2. bert – эксперименты с моделью bert

3. roberta – эксперименты с моделью roberta
4. albert – эксперименты с моделью albert
5. electra – эксперименты с моделью electra
6. distilbert – эксперименты с моделью distilbert
7. mobilebert – эксперименты с моделью mobilebert
8. labse – эксперименты с моделью labse
9. talkheads_ggelu_bert – эксперименты с моделью talkheads_ggelu_bert
10. lambert – эксперименты с моделью lambert
11. tn_bert – эксперименты с моделью tn_bert

В каждой папке выложено несколько файлов jupyter notebook, написан readme файл с указанием того, в каком файле что лежит, а также приведены ссылки на мой Yandex Disk для дообученных моделей.

Результаты экспериментов, их визуализация, анализ и выводы о сходстве датасетов:

1. для первой серии экспериментов (определение позитивности отзыва) – в файле results_of_first_series_of_experiments.ipynb
2. для второй серии экспериментов (определение спама) – в файле results_of_second_series_of_experiments.ipynb
3. для датасетов, взятых из разных серий экспериментов – в файле additional_research_for_similarity_in_second_sense_for_datasets_from_different_series_of_experiments.ipynb

Гипотезы, а также разработанные в соответствии с ними методы оценки сходства датасетов, исследование эффективности этих методов, выводы:

1. метод в соответствии с первой гипотезой – в файле first_method_for_comparing_datasets.ipynb
2. метод в соответствии со второй гипотезой – в файле second_method_for_comparing_datasets.ipynb

Практическая часть 5 – Краткое описание содержимого результатов экспериментов

В качестве результатов экспериментов для каждой модели выступают:

1) качество на обучающей выборке (для каждого из датасетов),
2) качество на тестовой выборке (для каждой пары вида *(датасет для обучения, датасет для тестирования)*),
3) изменение качества при переходе с тестовой выборки на обучающую (для каждой пары вида *(датасет для обучения, датасет для тестирования)*),

4) относительное (выраженное в процентах) изменение качества (accuracy, f1_score) при переходе с обучающей выборки на тестовую (для каждой пары вида *(датасет для обучения, датасет для тестирования)*),

После проведения экспериментов для двух датасетов A и B следует сохранять результаты таким образом, чтобы впоследствии их было удобно обрабатывать. Поэтому я записываю результаты в DataFrame и затем сохраняю его с расширением csv в свой Google Drive в соответствующую папку.

Так как результатов экспериментов получилось достаточно много, потребовалась их визуализация. Визуализация результатов экспериментов происходит в отдельном ноутбуке. Благодаря визуализации должно стать понятно, насколько схожими являются рассматриваемые датасеты; это должно помочь сформулировать гипотезы относительно того, на что опирается метод оценки сходства датасетов.

Практическая часть 6 – Результаты обучения и тестирования для первой серии экспериментов

Результаты первой и второй серии экспериментов выложены в гитхаб-репозиторий. Выполнена визуализация результатов. Наиболее значимые графики и выводы из них представлены ниже в отчете.

Рисунок 1. График разностей качества (accuracy, f1_score) на тестовой и обучающей выборке для 10 моделей

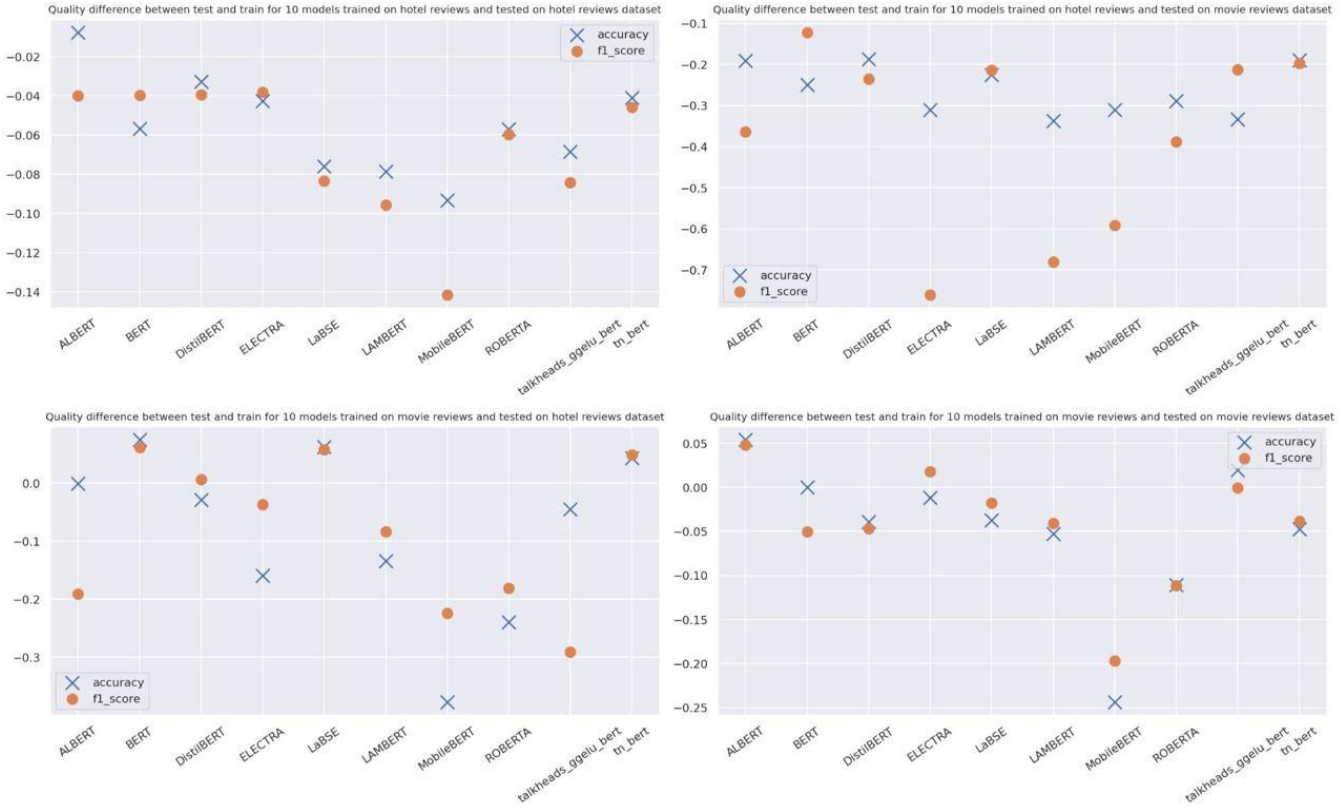


Рисунок 2. Разности качества на тестовой и обучающей выборках для всех 10 моделей сразу, а также среднее и медианное значения для разностей accuracy, precision, recall, f1_score.

Рисунок 3. График относительного изменения качества (accuracy, f1_score) при переходе с обучающей выборки на тестовую для каждой из 10 моделей

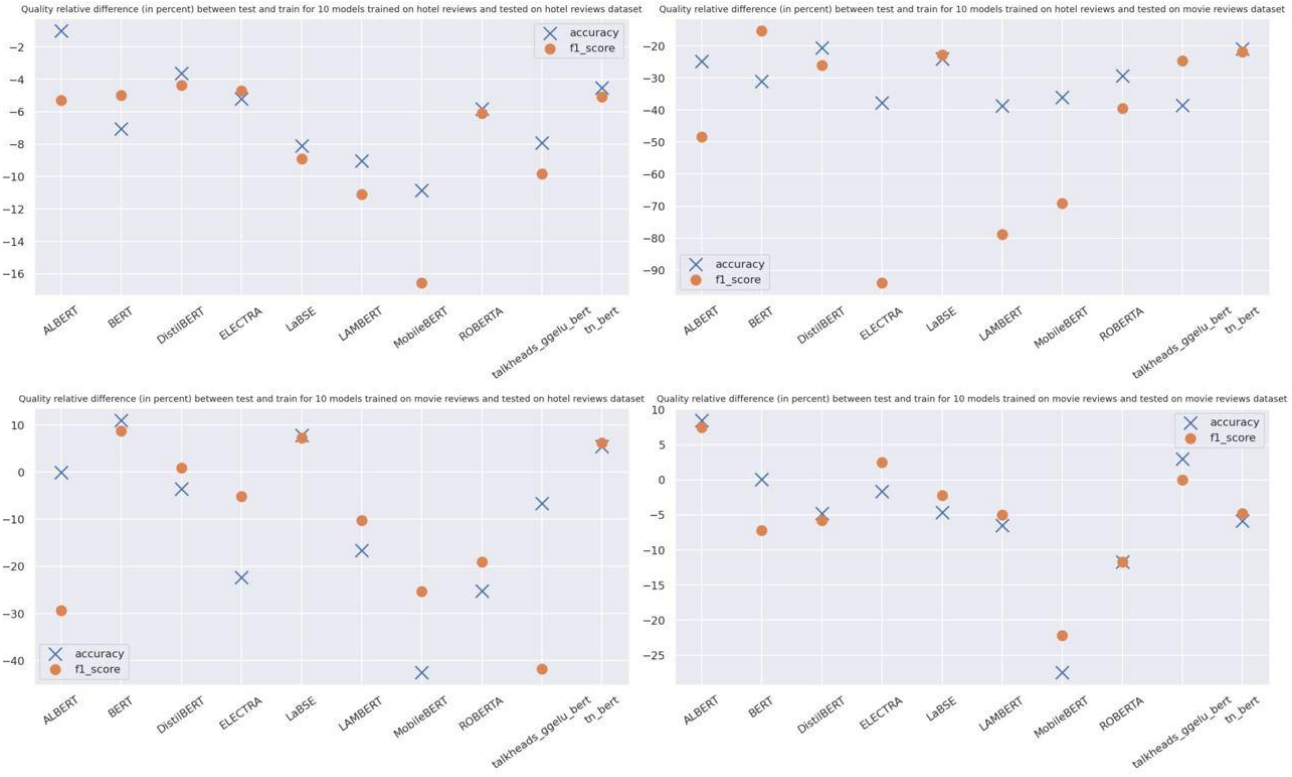
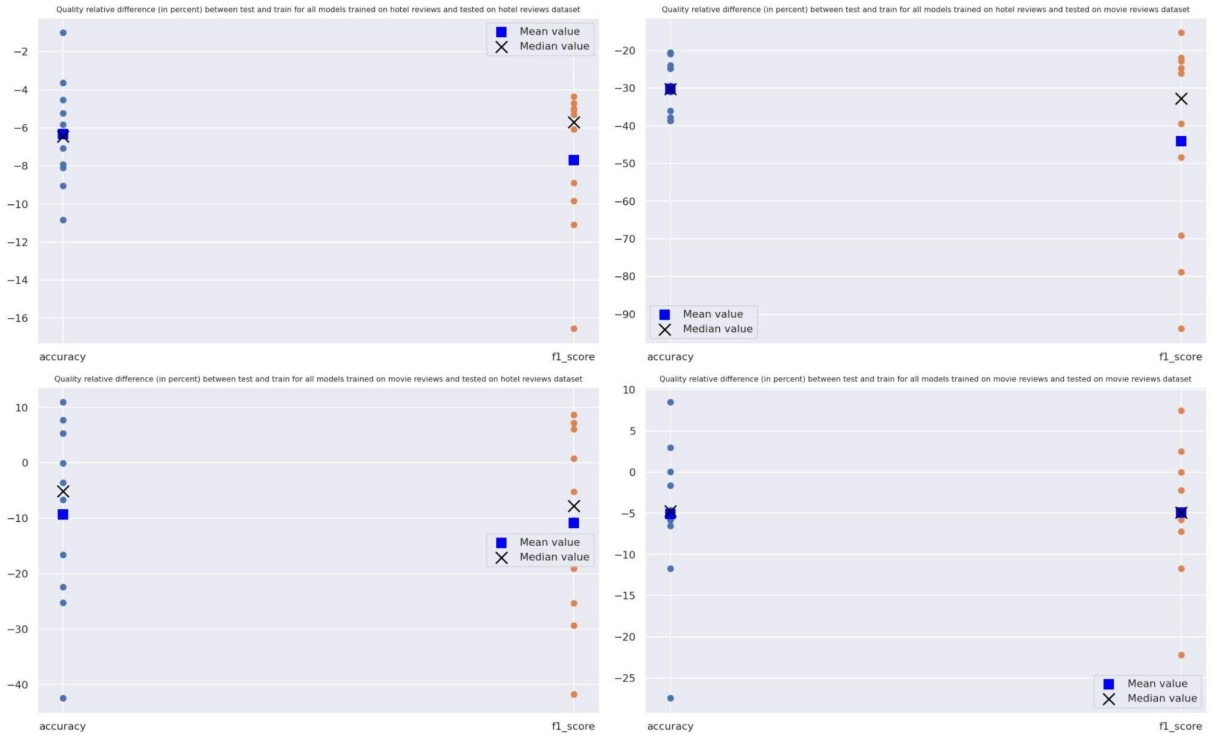


Рисунок 4. Относительные изменения качества (в процентах) на тестовой и обучающей выборках для всех 10 моделей сразу, а также среднее и медианное значения для относительных изменений accuracy, f1_score.



Первая серия экспериментов проводилась с балансировкой классов и в обучающей выборке, и в тестовой выборке. Позитивных и негативных отзывов было примерно одинаковое количество. Поэтому при анализе следует обращать внимание и на f1_score, и на accuracy.

Про сходство датасетов в первом смысле

По введенному определению сходства датасетов в первом смысле, модель, обученная на датасете A и показывающая хорошие результаты при тестировании на A, также показывает хорошие результаты при тестировании и на B. Следовательно, нужно проанализировать графики слева-внизу и справа-вверху.

Анализ графиков:

1) при обучении на датасете с отзывами о фильмах и тестировании на датасете с отзывами об отелях (переход с отзывов о фильмах на отзывы об отелях):

I. качество на тестовой выборке:

Значения accuracy при тестировании в зависимости от модели колеблются примерно от 0.51 до 0.87, среднее арифметическое и медианное значения в районе 0.7.

Значения f1_score при тестировании колеблются примерно от 0.4 до 0.87; среднее арифметическое значение f1_score примерно 0.7, тогда как медианное

значение f1_score в районе 0.75. У моделей ALBERT и talkheads_ggelu_bert наблюдаются наименьшие значения f1_score: примерно 0.4 и 0.46 соответственно. Для остальных моделей f1_score не менее 0.65.

II. абсолютное изменение качества:

При переходе с тестовой выборки на обучающую, изменение accuracy составляет от -0.38 до +0.08; среднее значение изменений accuracy составляет примерно -0.08, тогда как медианное – приблизительно -0.04. Только у двух моделей из десяти (MobileBERT и ROBERTA) изменение accuracy менее -0.18.

При переходе с обучающей выборки на тестовую, изменение f1_score, в зависимости от используемой модели, составляет от -0.3 до +0.07; среднее значение изменений f1_score составляет примерно -0.09, тогда как медианное – приблизительно -0.07.

III. относительное изменение качества:

При переходе с обучающей выборки на тестовую, у 9 моделей из 10 относительное изменение accuracy составляет от -26% до +11%, среднее значение по всем 10 моделям примерно -9%, а медианное значение – около -6%. Лишь у одной модели (MobileBERT) accuracy упало достаточно сильно, приблизительно на 42%.

При переходе с обучающей выборки на тестовую, у 9 моделей из 10 относительное изменение f1_score составляет от -29% до +9%, среднее значение по всем 10 моделям примерно -10%, а медианное значение – около -8%. Лишь у одной модели (talkheads_ggelu_bert) f1_score упало достаточно сильно, приблизительно на 41%.

2) при обучении на датасете с отзывами об отелях и тестировании на датасете с отзывами о фильмах (переход с отзывов об отелях на отзывы о фильмах):

I. качество на тестовой выборке:

Значения accuracy при тестировании в зависимости от модели колеблются примерно от 0.5 до 0.7, среднее арифметическое около 0.61, медианное значение приблизительно 0.58.

Значения f1_score при тестировании имеет значительный разброс, находясь в промежутке примерно от 0.05 до 0.73. Среднее значение f1_score составляет около 0.5, тогда как медианное – примерно 0.62.

II. абсолютное изменение качества:

При переходе с обучающей выборки на тестовую, изменение accuracy, в зависимости от используемой модели, составляет от -0.34 до -0.19; среднее и медианное значения изменений accuracy составляют примерно -0.27.

Изменение f1_score при переходе с обучающей выборки на тестовую имеет большой разброс. Изменения f1_score составляют от -0.77 до -0.12. Среднее значение примерно -0.38, тогда как медианное -0.3.

III. относительное изменение качества:

При переходе с обучающей выборки на тестовую, относительное изменение accuracy составляет от -40% до -20%, среднее и медианное значения примерно -30%.

Относительное изменение f1_score при переходе с обучающей выборки на тестовую имеет большой разброс. Относительное изменение f1_score (в зависимости от модели) составляет от -95% до -15%. Среднее арифметическое значение примерно -43%, тогда как медианное – приблизительно -33%.

Заключение по сходству датасетов в первом смысле

В большинстве своём, модели, обученные на датасете с отзывами о фильмах, также показывают хорошие результаты при тестировании на датасете с отзывами об отелях. Таким образом, наблюдается перенос качества с датасета с отзывами о фильмах на датасет с отзывами об отелях, то есть имеет место сходство датасетов в первом смысле.

При этом, перенос качества в обратную сторону (с датасета с отзывами об отелях на датасет с отзывами о фильмах) происходит заметно хуже. Убедились, что сходство датасетов в первом смысле не обладает свойством симметричности.

Про сходство датасетов во втором смысле

По введенному нами определению сходства датасетов во втором смысле, датасеты A и B схожи во втором смысле, если множество моделей, хорошо работающих на A (при обучении на A), совпадает с множеством моделей, хорошо работающих на B (при обучении на B).

1) Для датасета с отзывами об отелях

I. качество на тестовой выборке:

Accuracy принимает значения от 0.75 до 0.95, среднее значение accuracy в районе 0.82, тогда как медианное – около 0.79.

f1_score принимает значения от 0.71 до 0.95, среднее значение f1_score в районе 0.8, тогда как медианное – около 0.77.

II. абсолютное изменение качества:

При переходе с обучающей выборки на тестовую значение accuracy изменяется не более чем на -0.1; среднее и медианное значения в районе -0.06.

Для 9 моделей из 10 изменение f1_score при переходе с обучающей выборки на тестовую принимает значения от -0.1 до -0.035; среднее и

медианное значения (по всем десяти моделям) соответственно равны примерно -0.065 и -0.05. Самое сильное падение наблюдалось у модели MobileBERT, где изменение f1_score составило примерно -0.14.

III. относительное изменение качества:

Относительное изменение accuracy при переходе с тестовой выборки на обучающую составляет от -11% до -1%; медианное и среднее значения в районе -6%.

Для 9 моделей из 10 относительное изменение f1_score при переходе с обучающей выборки на тестовую составляет от -11% до -4%. Среднее значение изменения f1_score (по всем 10 моделям) около -8%, тогда как медианное приблизительно -6%. Самое сильное относительное изменение f1_score наблюдалось у модели MobileBERT и составило примерно -0.16%.

2) Для датасета с отзывами о фильмах

I. качество на тестовой выборке:

Accuracy принимает значения от 0.64 до 0.84, среднее и медианное значения accuracy в районе 0.73.

f1_score принимает значения от 0.65 до 0.84, среднее и медианное значения f1_score около 0.74.

II. абсолютное изменение качества:

Для 9 моделей из 10 изменение accuracy при переходе с обучающей выборки на тестовую принимает значения от -0.11 до +0.05; среднее и медианное значения (по всем десяти моделям) приблизительно равны -0.05. Самое сильное падение наблюдалось у модели MobileBERT, где изменение accuracy составило примерно -0.25.

Интересно, что для f1_score наблюдается почти то же самое. У 9 моделей из 10 изменение f1_score при переходе с обучающей выборки на тестовую составляет от -0.11 до +0.05. Среднее и медианное значения (по всем десяти моделям) приблизительно равны -0.05. Самое сильное изменение f1_score наблюдалось у модели MobileBERT и составило примерно -0.2.

III. относительное изменение качества:

Для 9 моделей из 10 относительное изменение accuracy при переходе с обучающей выборки на тестовую принимает значения от -12% до +8%; среднее и медианное значения (по всем десяти моделям) приблизительно равны -5%. Самое сильное падение наблюдалось у модели MobileBERT, где изменение accuracy составило примерно -28%.

У 9 моделей из 10 относительное изменение f1_score при переходе с обучающей выборки на тестовую составляет от -12% до +7%. Среднее и медианное значения (по всем десяти моделям) приблизительно равны -5%.

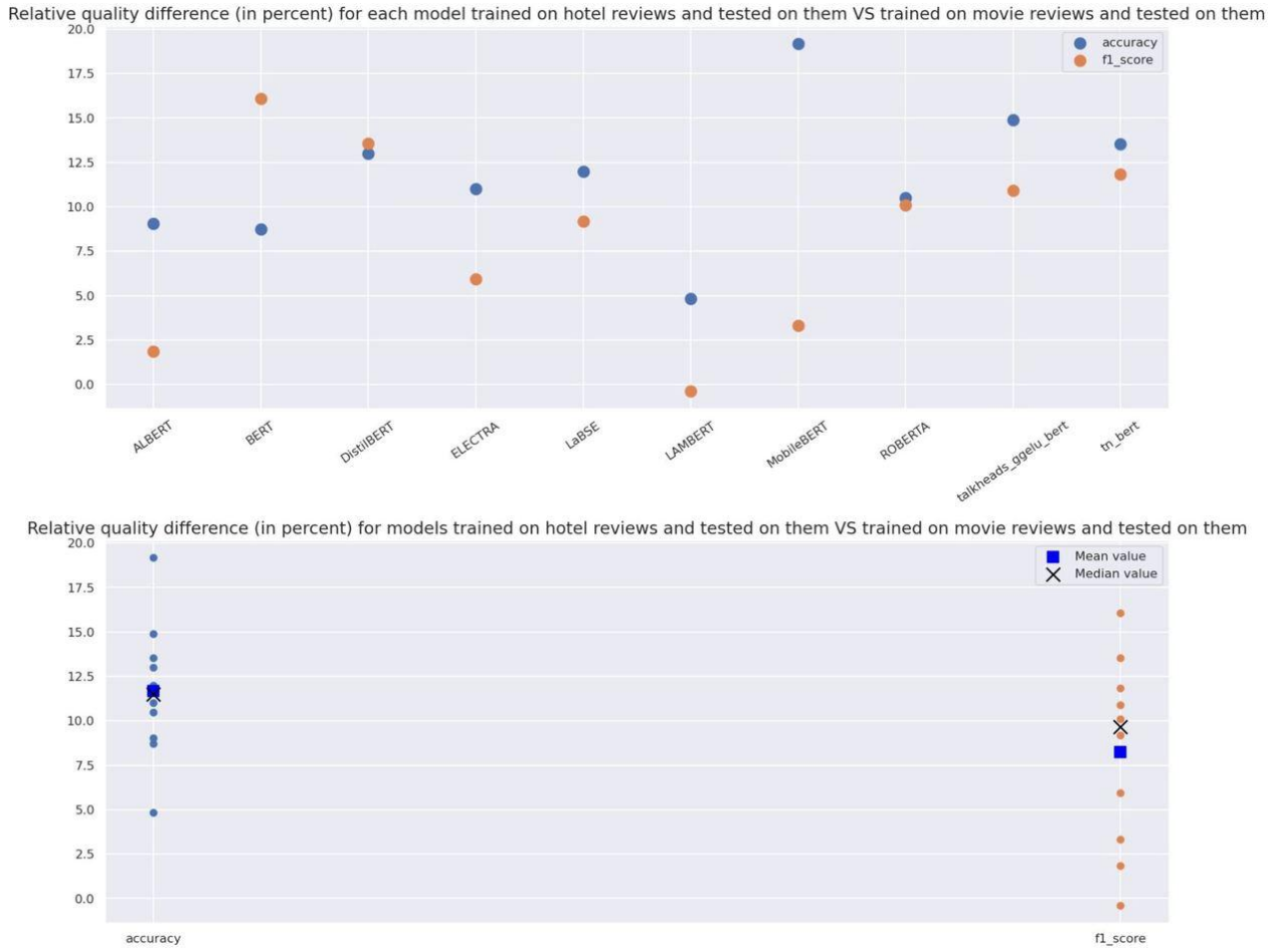
Самое сильное относительное изменение f1_score наблюдалось у модели MobileBERT и составило примерно -22%.

Заключение по сходству датасетов во втором смысле

Можно утверждать, что в большинстве своем модели хорошо работают и на датасете с отзывами об отелях, и на датасете с отзывами о фильмах. При этом у модели MobileBERT при переходе с обучающей выборки на тестовую качество падает сильнее, чем у других моделей. Это происходит на каждом из двух рассматриваемых датасетов. Подводя итог, скорее всего **наблюдается сходство датасетов во втором смысле**.

Рисунок 5. Абсолютное изменение качества для каждой модели при переходе с датасета movie reviews на датасет hotel reviews (подробно описано выше)

Рисунок 6. Относительное изменение качества для каждой модели при переходе с датасета movie reviews на датасет hotel reviews, выраженное в процентах



Анализ:

Переходом с датасета A на датасет B будем называть переход

от обучения и тестирования на датасете A

к обучению и тестированию на датасете B.

1) Абсолютное изменение accuracy при переходе с датасета movie reviews на датасет hotel reviews составляет от 0.035 до 0.125 (в зависимости от модели). Среднее значение примерно 0.085, медианное значение приблизительно 0.09.

2) Абсолютное изменение f1_score при переходе с датасета movie reviews на датасет hotel reviews составляет от -0.005 до 0.105. Среднее значение примерно 0.06, медианное значение приблизительно 0.075.

3) Относительное изменение accuracy при переходе с датасета movie reviews на датасет hotel reviews составляет от 5% до 19%. Среднее и медианное значения приблизительно 11.5%.

4) Относительное изменение f1_score при переходе с датасета movie reviews на датасет hotel reviews составляет от -0.5% до 16%. Среднее значение примерно 8%, тогда как медианное значение приблизительно 9.5%.

Строго говоря, изменение не является стабильным, оно принимает разные значения в зависимости от рассматриваемой модели.

Тем не менее, и абсолютное, и относительное изменение качества (accuracy и f1_score) при переходе с датасета movie reviews на датасет hotel reviews небольшое. **Это подтверждает сходство датасетов во втором смысле.**

Итоговый вывод по первой серии экспериментов:

Наблюдается сходство датасетов и в первом, и во втором смысле.

Практическая часть 7 – Результаты обучения и тестирования для второй серии экспериментов

Рисунок 7. График разностей качества (accuracy, f1_score) на тестовой и обучающей выборке для 10 моделей

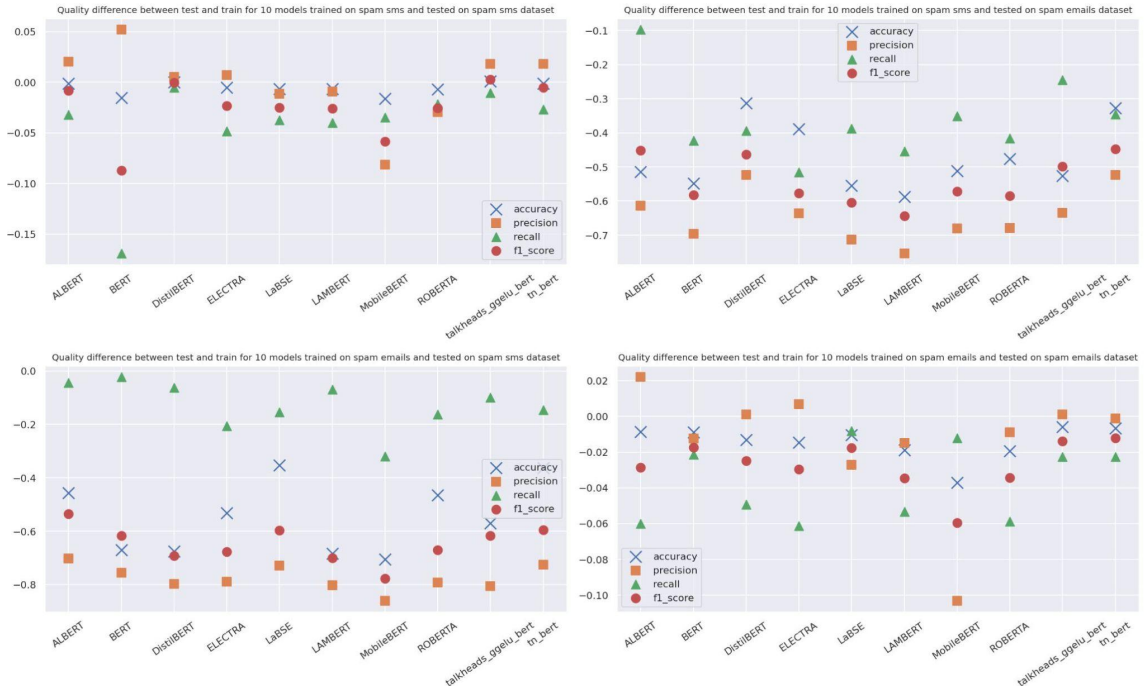


Рисунок 8. Разности качества на тестовой и обучающей выборках для всех 10 моделей сразу, а также среднее и медианное значения для разностей accuracy, precision, recall, f1_score.

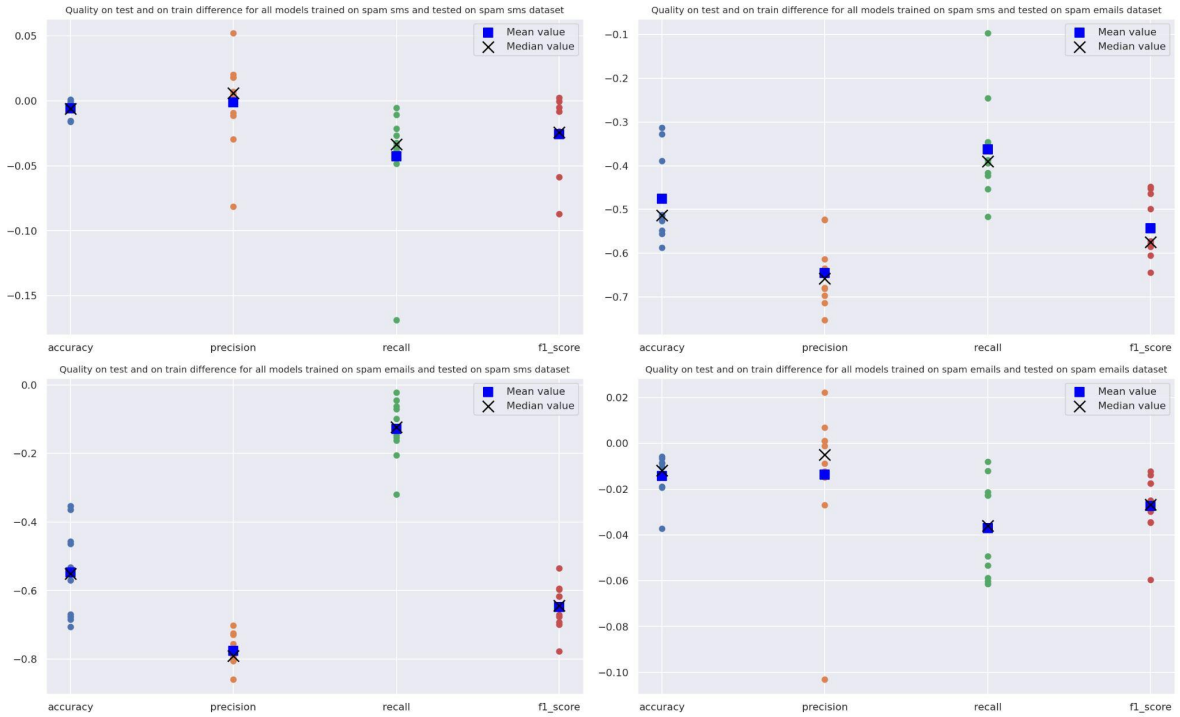


Рисунок 9. График относительного изменения качества (accuracy, f1_score) при переходе с обучающей выборки на тестовую для каждой из 10 моделей. Относительное изменение качества выражено в процентах

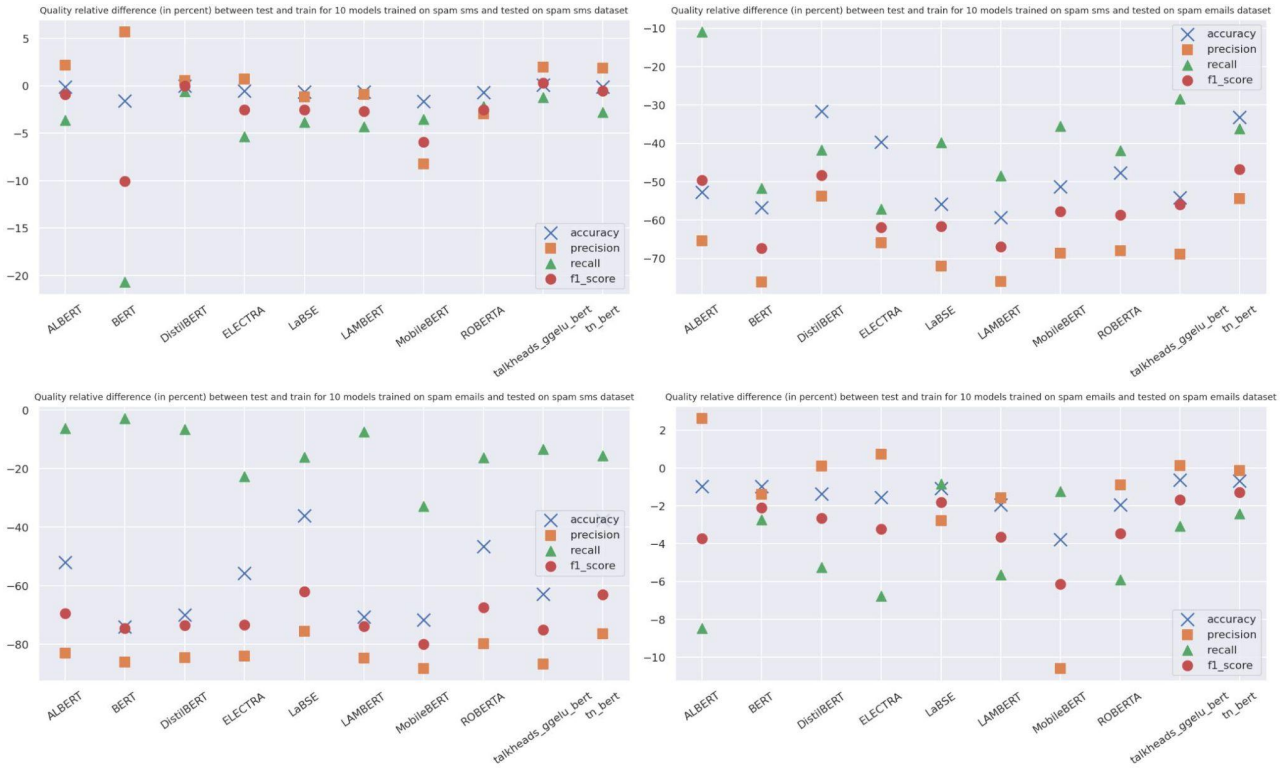
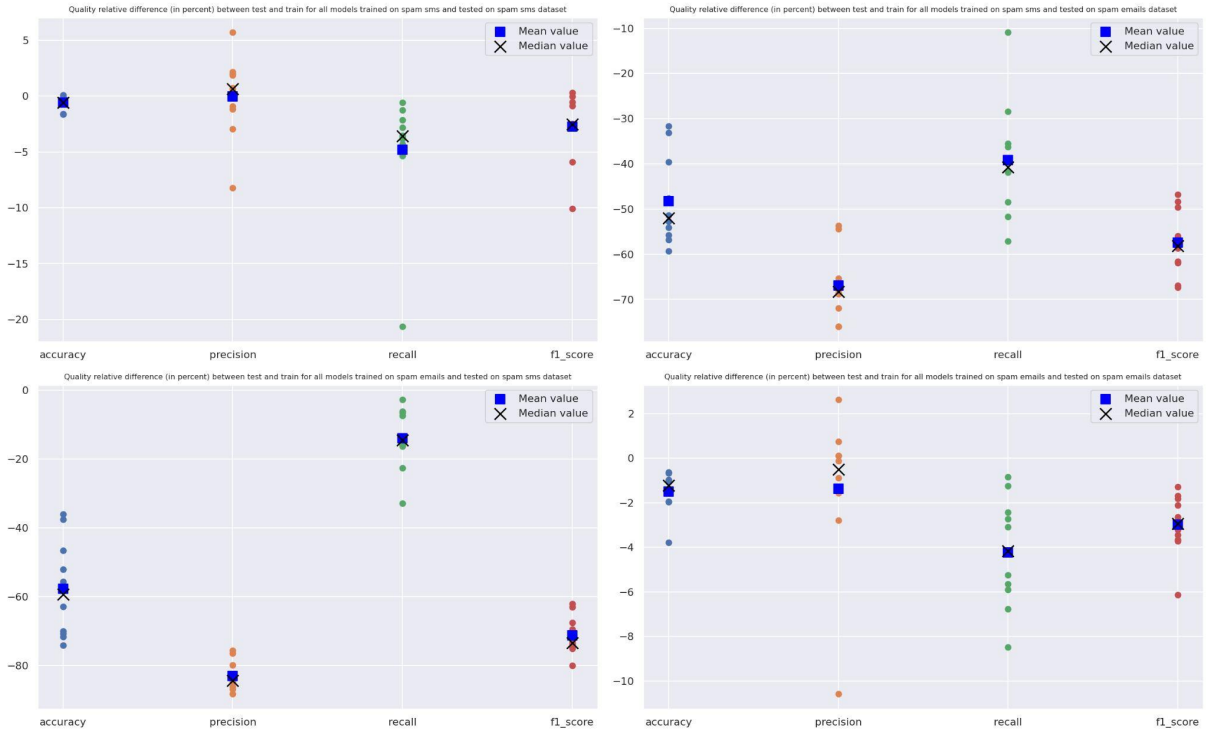


Рисунок 10. Относительные изменения качества (в процентах) на тестовой и обучающей выборках для всех 10 моделей сразу, а также среднее и медианное значения для относительных изменений accuracy, f1_score



Вторая серия экспериментов, в отличие от первой, проводилась без балансировки классов в обучающей и тестовой выборках. Сообщений или электронных писем, которые помечены как спам, заметно меньше. Поэтому при анализе второй серии экспериментов следует обращать внимание не на accuracy и f1_score (как мы делали при анализе результатов первой серии экспериментов), а исключительно на f1_score.

Во второй серии экспериментов датасеты разбивались на две примерно равные части: обучающую выборку и тестовую выборку.

Про сходство датасетов в первом смысле

По введенному определению сходства датасетов в первом смысле, модель, обученная на датасете A и показывающая хорошие результаты при тестировании на A, также показывает хорошие результаты при тестировании и на B. Значит, нужно проанализировать графики слева-внизу и справа-вверху.

Анализ графиков:

1) при обучении на датасете с sms и тестировании на датасете с emails (переход с sms на emails):

I. качество на тестовой выборке:

Значения f1_score при тестировании в зависимости от модели колеблются примерно от 0.3 до 0.5; среднее и медианное значения f1_score в районе 0.4.

II. абсолютное изменение качества:

f1_score, в зависимости от используемой модели, уменьшается на примерно 0.45–0.65; среднее и медианное значения составляют примерно 0.55 и 0.58 соответственно.

III. относительное изменение качества:

Относительное падение f1_score составляет от 46% до 68%, среднее и медианное значения в районе 58%.

2) при обучении на датасете с emails и тестировании на датасете с sms (переход с emails на sms):

I. качество на тестовой выборке:

Значения f1_score при тестировании в зависимости от модели колеблются примерно от 0.2 до 0.38; среднее и медианное значения f1_score в районе 0.26.

II. абсолютное изменение качества:

f1_score, в зависимости от используемой модели, уменьшается на примерно 0.46–0.78; среднее и медианное значения составляют примерно 0.64.

III. относительное изменение качества:

Относительное падение f1_score составляет от 61% до 80%, среднее и медианное значения в районе 71-73%.

Заключение по сходству датасетов в первом смысле:

При переходе с sms на emails значения f1_score получаются выше, чем при переходе с emails на sms. Также при переходе с sms на emails падение f1_score ниже.

Таким образом, перенос качества с датасета с sms на датасет с emails происходит лучше, чем перенос качества наоборот (с датасета с emails на датасет с sms). Следовательно, перенос качества несимметричен.

Вместе с тем, даже при наиболее удачном из двух рассматриваемых переносов среднее и медианное значения f1_score составляют примерно 0.4, что не соответствует высокому качеству.

Подводя итог, **не наблюдается сходство датасетов в первом смысле (нет переноса качества)**.

Про сходство датасетов во втором смысле

1) Для датасета с sms

I. качество на тестовой выборке:

9 из 10 моделей имеют f1_score примерно от 0.89 до 0.98, медианное и среднее значения в районе 0.92-0.94. Модель Bert имеет f1_score около 0.78, что существенно ниже, чем у остальных моделей.

II. абсолютное изменение качества:

Изменение f1_score при переходе с тестовой выборки на обучающую принимает значения от -0.09 до 0.01; среднее и медианное значения в районе -0.025.

III. относительное изменение качества:

Относительное изменение f1_score при переходе с тестовой выборки на обучающую составляет от -10% до 1%; медианное и среднее значения в районе -3%.

2) Для датасета с emails

I. качество на тестовой выборке:

9 из 10 моделей имеют f1_score примерно от 0.80 до 0.96, медианное и среднее значения примерно 0.91 и 0.88 соответственно. Модель ALBERT имеет f1_score около 0.74, что существенно ниже, чем у остальных моделей.

II. абсолютное изменение качества:

Изменение f1_score при переходе с тестовой выборки на обучающую принимает значения от -0.06 до -0.01; среднее и медианное значения в районе -0.0275.

III. относительное изменение качества:

Относительное изменение f1_score при переходе с тестовой выборки на обучающую составляет от -6% до -1%; медианное и среднее значения в районе -3%.

Заключение по сходству датасетов во втором смысле:

Можно утверждать, что в большинстве своем модели хорошо работают и на датасете с sms, и на датасете с emails. Модель BERT показала не такой хороший результат, как остальные модели, на датасете с sms (f1_score = 0.78). А модель ALBERT показала не такой хороший результат на датасете с emails (f1_score около 0.74). Подводя итог, **наблюдается сходство датасетов во втором смысле.**

Рисунок 11. Абсолютное изменение качества для каждой модели при переходе с датасета spam emails на датасет spam sms

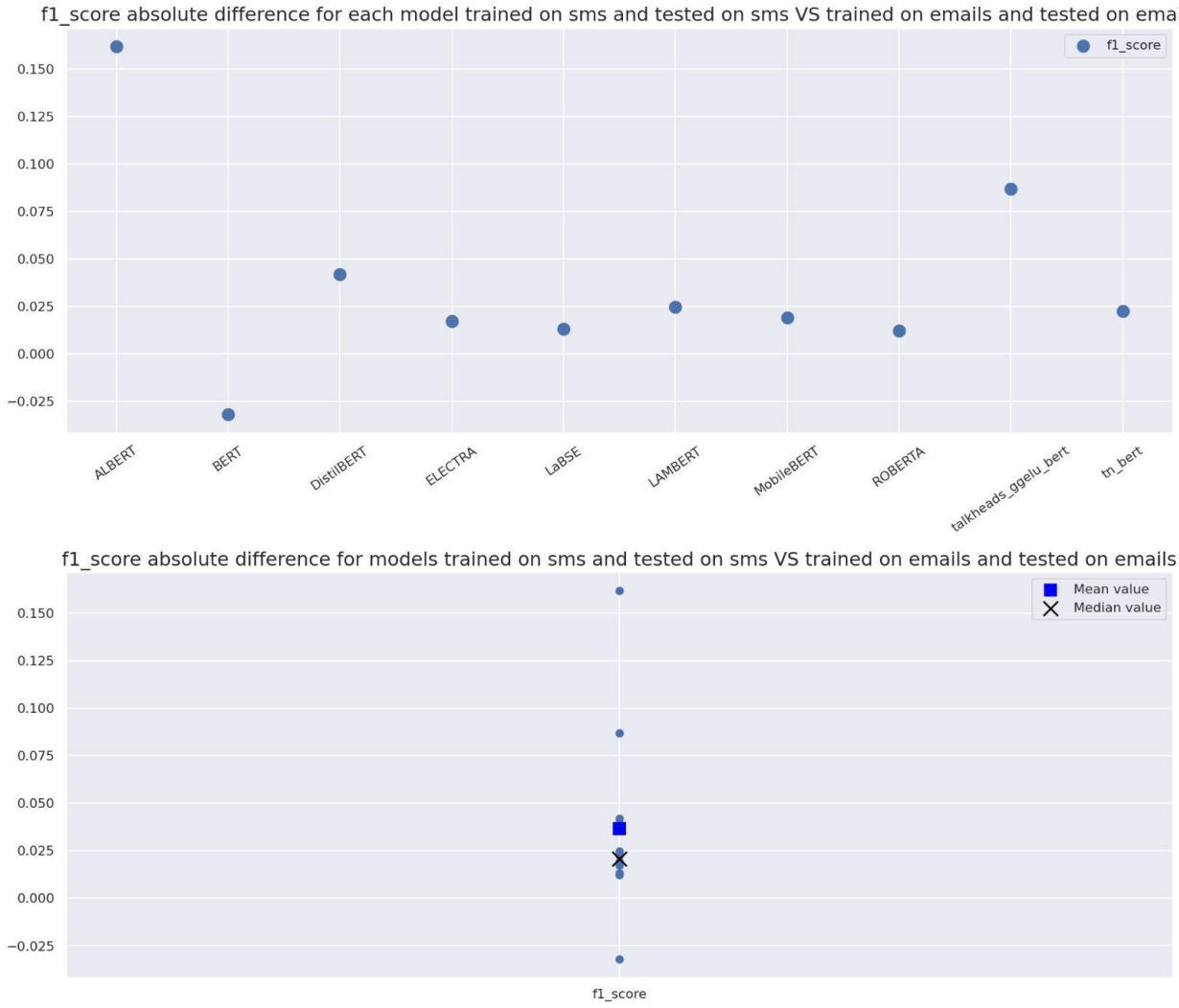
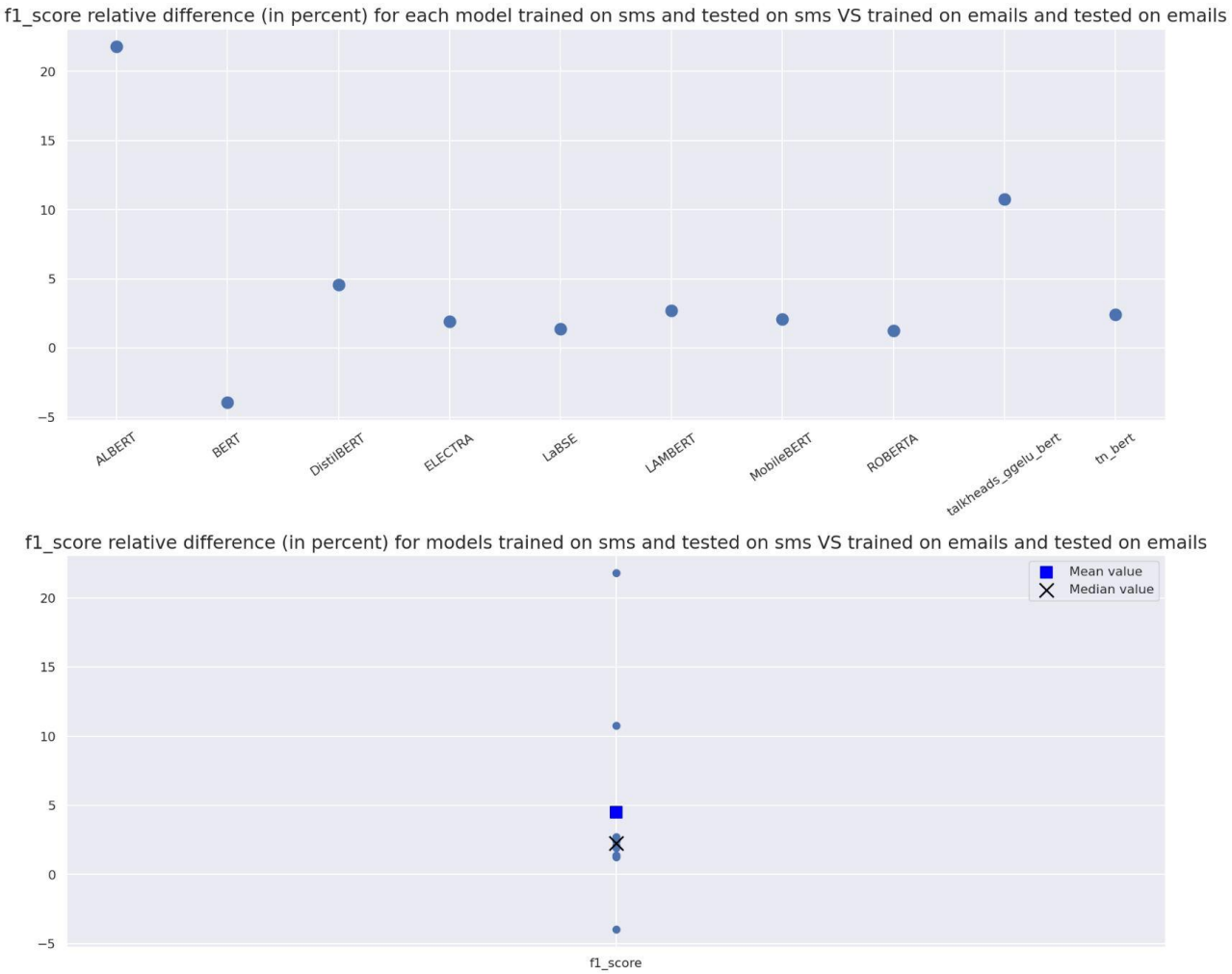


Рисунок 12. Относительное изменение качества для каждой модели при переходе с датасета spam emails на датасет spam sms, выраженное в процентах



Анализ:

Переходом с датасета A на датасет B будем называть переход

от обучения и тестирования на датасете A

к обучению и тестированию на датасете B.

1) Абсолютное изменение f1_score при переходе с датасета spam emails на датасет spam sms составляет от -0.03 до 0.16 (в зависимости от модели). Среднее значение примерно 0.04, медианное значение приблизительно 0.02.

Интересно, что для 7 моделей из 10 абсолютное изменение f1_score почти одинаковое, его можно назвать стабильным: оно находится в промежутке от 0.01 до 0.04.

У двух моделей (ALBERT и talkheads_ggelu_bert) абсолютное изменение выше, чем у 7 моделей из предыдущего абзаца, а у одной модели (BERT) – меньше.

2) Относительное изменение f1_score при переходе с датасета spam emails на датасет spam sms составляет от -4% до 22%. Среднее значение чуть меньше 5%, тогда как медианное значение приблизительно 2.5%.

Следует отметить, что для тех же самых 7 моделей из 10 относительное изменение f1_score можно назвать стабильным: оно находится в промежутке от 1% до 5%.

Таким образом, и абсолютное, и относительное изменение качества (f1_score) при переходе с датасета spam emails на датасет spam sms небольшое. Для 7 моделей из 10 изменение почти одинаковое, его можно назвать стабильным. Всё это **подтверждает сходство датасетов во втором смысле**.

Итоговый вывод по второй серии экспериментов:

Наблюдается сходство датасетов только во втором смысле.

Экспериментальная часть 1 – Гипотезы для метода оценки сходства датасетов

Исходя из результатов практической части сформулировано несколько гипотез относительно того, в чем может проявляться сходство датасетов.

1) сходство датасетов в первом или втором смысле может быть связано со схожестью их текстового содержания, которое отражается в схожести часто встречающихся в датасетах слов;
2) рассматриваемые в практической части языковые модели берут от текстов эмбеддинги, поэтому сходство датасетов в первом или втором смысле может быть связано со сходством усредненного значения эмбеддингов (для каждого датасета своё усредненное значение эмбеддинга).

Экспериментальная часть 2 – Первый метод оценки сходства датасетов

Чтобы проверить первую гипотезу, реализуем метод оценки сходства датасетов, который заключается в следующем:

Посчитать наиболее часто встречающиеся слова (исключив предлоги и артикли, которые могут встречаться часто, но при этом не оказывать существенного влияния на смысл текста) в каждом из датасетов и оставить, например, топ из num:=100 наиболее часто встречающихся слов. Для каждого датасета этот топ будет свой. Затем оценить, насколько похожи эти топы слов: для каждого слова из первого топа найти максимум similarity(рассматриваемое_фиксированное_слово_из_первого_топа, слово_из_второго_топа) по всем возможным значениям переменной слово_из_второго_топа. Проще говоря, для каждого слова из первого топа найти наиболее схожее слово из второго топа; после чего запомнить число, соответствующее степени их схожести.

Для каждого слова из первого топа получится какое-то значение максимума similarity. Следует сложить эти значения. Чем больше полученная сумма, тем более вероятно сходство датасетов.

Разумеется, можно брать num равным не 100, а, например, 200 или 1000. Так как не ясно, какое значение num даст наилучшие результаты, то следует провести вычисления для различных значений num.

Заметим, что этот метод не является симметричным, то есть, в общем случае, результат может измениться, если передать методу те же самые датасеты *в другом порядке*.

Экспериментальная часть 3 – Второй метод оценки сходства датасетов

Чтобы проверить вторую гипотезу, также был разработан метод оценки сходства датасетов:

Сделать векторные представления от каждого текста из датасета (например, с помощью bert) и усреднить. Для разных датасетов сравнить полученный эмбеддинг, например, с помощью cosine_similarity. Чем больше полученное значение, тем более похожими являются полученные усредненные эмбеддинги и, по нашей гипотезе, тем больше сходство сравниваемых датасетов.

Следует отметить, что метод является симметричным, то есть, если передать методу те же самые датасеты в другом порядке, возвращаемое им численное значение **не изменится**. Таким образом, этот метод имеет ограниченную ценность вот в каком смысле: он, скорее всего, не будет давать хороших результатов при оценке сходства датасетов во втором смысле, так как сходство датасетов во втором смысле **зависит** от того, с какого датасета на какой мы переходим. Тем не менее, данный метод может быть полезен для предсказания сходства датасетов в первом смысле.

Экспериментальная часть 4 – Методика исследования эффективности метода

Напомним, метод оценки сходства датасетов представляет собой функцию, которая в качестве аргументов принимает два датасета, а возвращает численное значение. Это численное значение должно быть тем больше, чем более вероятно сходство датасетов.

Методика исследования эффективности разработанных методов для сходства двух датасетов в первом смысле:

Рассматривая различные пары датасетов, посчитаем корреляцию между численными значениями, которые возвращает метод сравнения датасетов, и относительным или абсолютным изменением качества при переходе с обучающей выборки первого датасета на тестовую выборку второго датасета. Для каждой из N=10 моделей изменение качества будет своё, и нужно учитывать их все. Поэтому мы усредним полученные значения или же возьмем медианное значение изменения качества при переходе с обучающей выборки первого датасета на тестовую выборку второго датасета.

Рассматривая различные пары датасетов, как уже было обосновано в практической части, сходство датасетов в первом смысле может наблюдаться только между теми датасетами, которые взяты из одной и той же серии

экспериментов. Для тех пар датасетов, где оба датасета принадлежат одной и той же серии экспериментов, изменение качества вычисляется так, как описано в абзаце выше. Допускается рассматривать пары, где в качестве первого и второго элементов пары представлен один и тот же датасет, поскольку такая пара удовлетворяет ограничению.

Мы также рассматриваем те пары датасетов, где датасеты взяты из разных серий экспериментов. Для них будем считать, что качество падает до нулевого (при переходе с обучающей выборки первого датасета на тестовую выборку второго датасета). Это полностью соответствует тому, что датасеты, взятые из разных серий экспериментов, не могут быть схожи в первом смысле.

Методика исследования эффективности разработанных методов для сходства двух датасетов во втором смысле:

Рассматривая различные пары датасетов, посчитаем корреляцию между численными значениями, которые возвращает метод сравнения датасетов, и относительным или абсолютным изменением качества при переходе с первого датасета на второй. Здесь переходом с датасета A на датасет B является переход от обучения и тестирования на датасете A к обучению и тестированию на датасете B. Разумеется, для каждой из N=10 моделей изменение качества будет своё, и нужно учитывать их все. Поэтому мы усредним полученные значения или же возьмем медианное значение изменения качества при переходе с одного датасета на другой.

Для сходства датасетов во втором смысле можно использовать все возможные пары различных датасетов.

Продолжение для обеих методик:

Чем выше получается корреляция, тем эффективнее разработанный метод оценки сходства датасетов. Разумеется, результат вычисления корреляции зависит от того, какой набор пар датасетов мы рассматриваем.

Есть несколько разумных вариантов составить набор рассматриваемых пар датасетов. При этом не очевидно, какой вариант является оптимальным. Например, не понятно, следует ли для сходства датасетов в первом смысле включать такие пары, где первый датасет совпадает со вторым. Поэтому рассмотрим два *случая*: с включением таких пар и без включения таких пар.

Более подробно про *случаи* (варианты составить набор рассматриваемых пар датасетов для вычисления корреляции) написано далее в экспериментальной части, в разделе “результаты исследования эффективности для первого (второго) метода”. Для каждого *случая* получается некоторое значение коэффициента корреляции. При оценке эффективности метода учитывается корреляция, получающаяся в **каждом** из *случаев*.

Замечание для первого метода:

Для первого метода стоит отметить: заранее не понятно, какое значение *num* (*num* – это число наиболее часто встречающихся слов) будет давать лучшие результаты. Я рассматриваю следующие значения num: 10, 20, 40, 80, 160, 320, 640, 900, 1280, 1800, 2560. Описанное выше исследование должно быть проведено для каждого из рассматриваемых значений *num*. Отмечу, что увеличивать num далее не имеет смысла, так как метод становится слишком затратным в плане вычислений.

Экспериментальная часть 5 – Критерии эффективности метода

Таблица 1. Описание корреляции

Значение коэффициента корреляции Пирсона	**Описание корреляции**
менее 0.05	отсутствие корреляции
от 0.05 до 0.3	очень слабая
от 0.3 до 0,5	слабая
от 0.5 до 0,7	средняя
от 0.7 до 0.9	высокая
более 0.9	очень высокая

Таблица 2. Критерии эффективности метода

Эффективность метода	**Корреляции для рассматриваемых двух случаев**
абсолютно неэффективный	обе корреляции не выше слабой
слабо эффективный	одна корреляция очень слабая, другая средняя, высокая или очень высокая
условно эффективный	обе корреляции средние; одна корреляция слабая, другая корреляция средняя, высокая или очень высокая
эффективный	одна корреляция средняя, другая высокая или очень высокая;

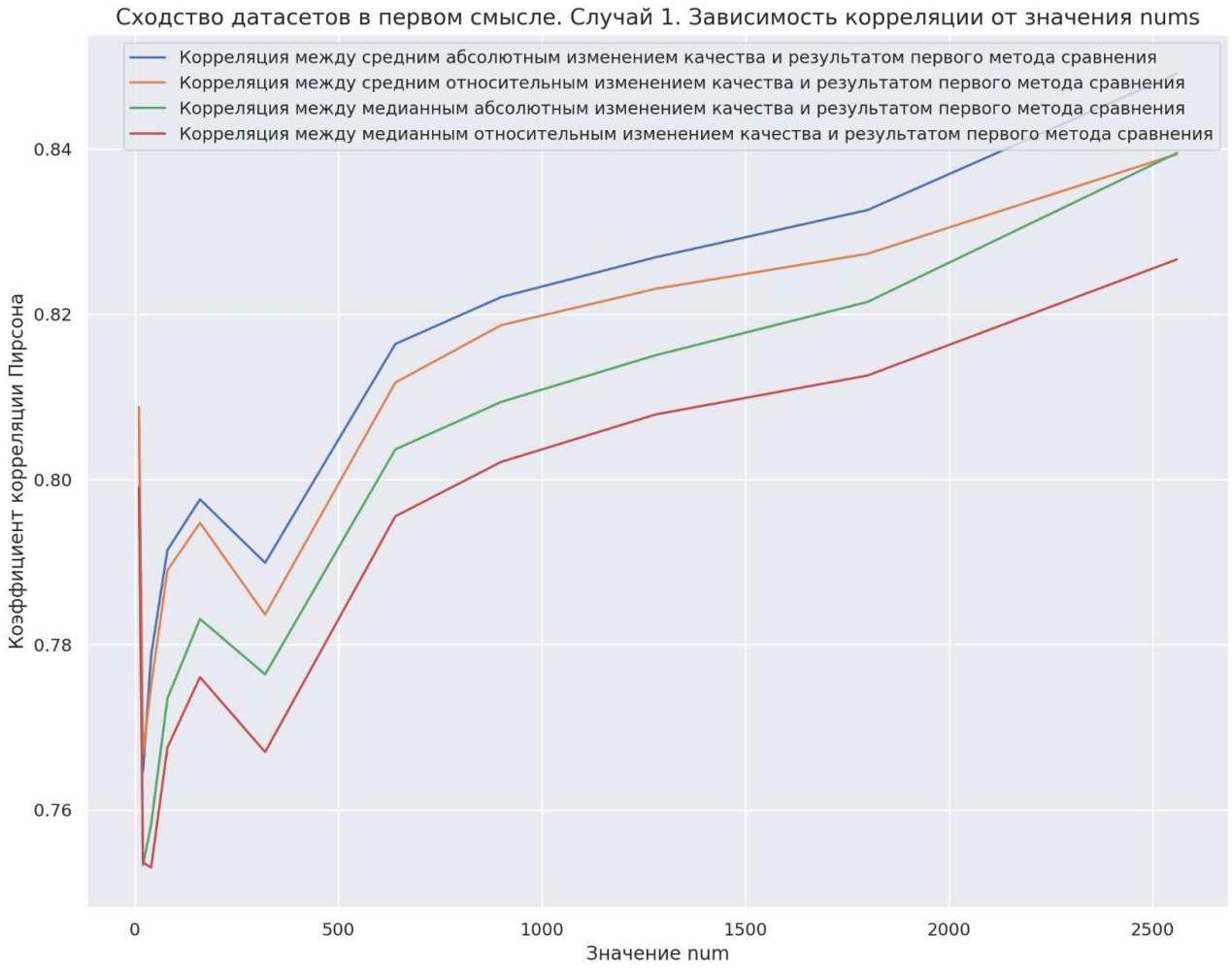
	одна корреляция высокая, другая высокая или очень высокая
абсолютно эффективный	обе корреляции очень высокие

Экспериментальная часть 6 – Результаты исследования эффективности для первого метода

Для сходства датасетов в первом смысле:

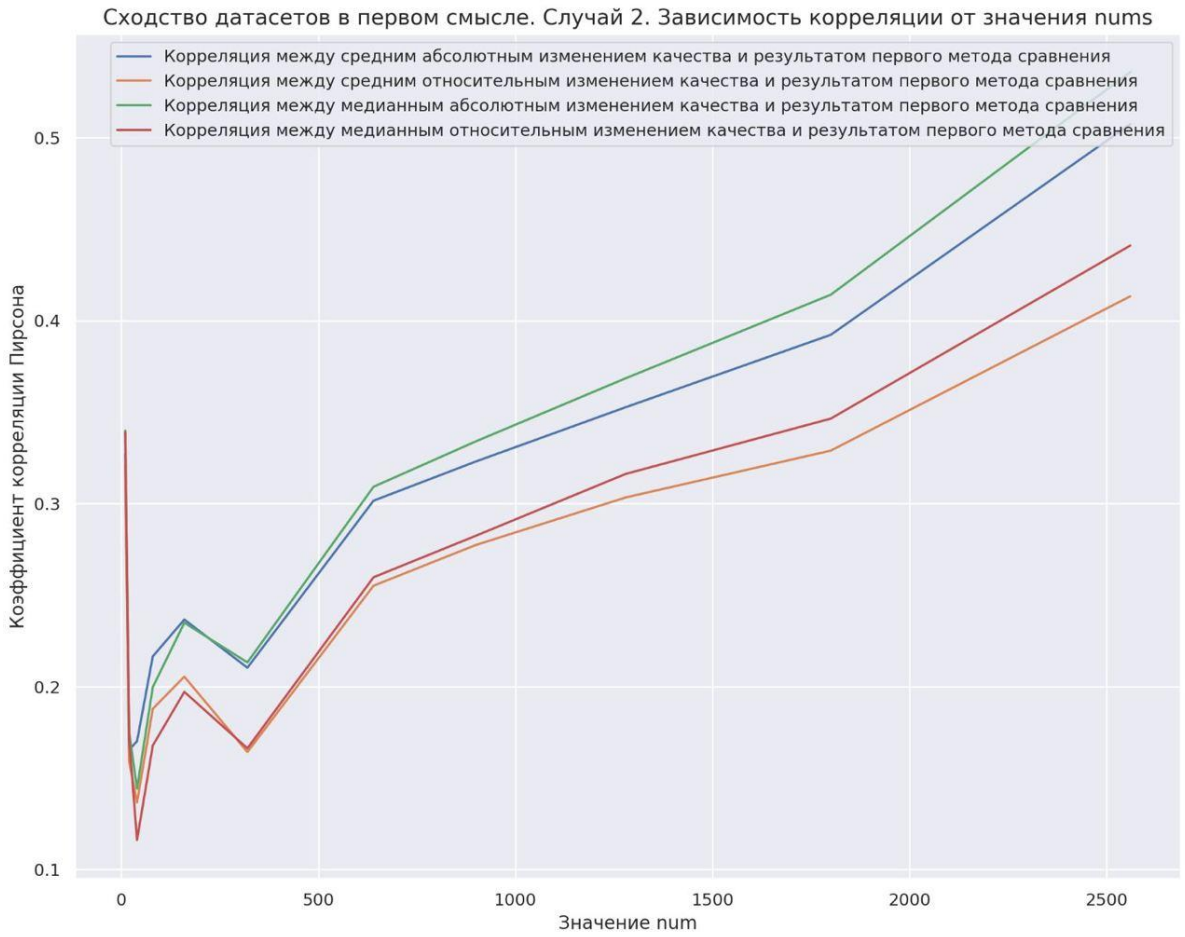
Случай 1: рассматриваются все возможные пары датасетов, в том числе пары с совпадающими датасетами.

Рисунок 13



Случай 2: рассматриваются все возможные пары **различных** датасетов.

Рисунок 14



Вывод:

Среднее и медианное изменение абсолютного качества, среднее и медианное изменение относительного качества – все они ведут себя схожим образом для обоих случаев (см. случаи 1 и 2 выше).

Для случая 1: Коэффициент корреляции, в зависимости от выбранного значения num и того, какое конкретно изменение качества рассматривается, принимает значения от 0.75 до 0.85, то есть в целом корреляция высокая. Сначала с увеличением значения num значение коэффициента корреляции несколько падает, потом немного колеблется, а затем растет. Оптимальным является num=2560, при котором значение коэффициента корреляции составляет от 0.82 до 0.85.

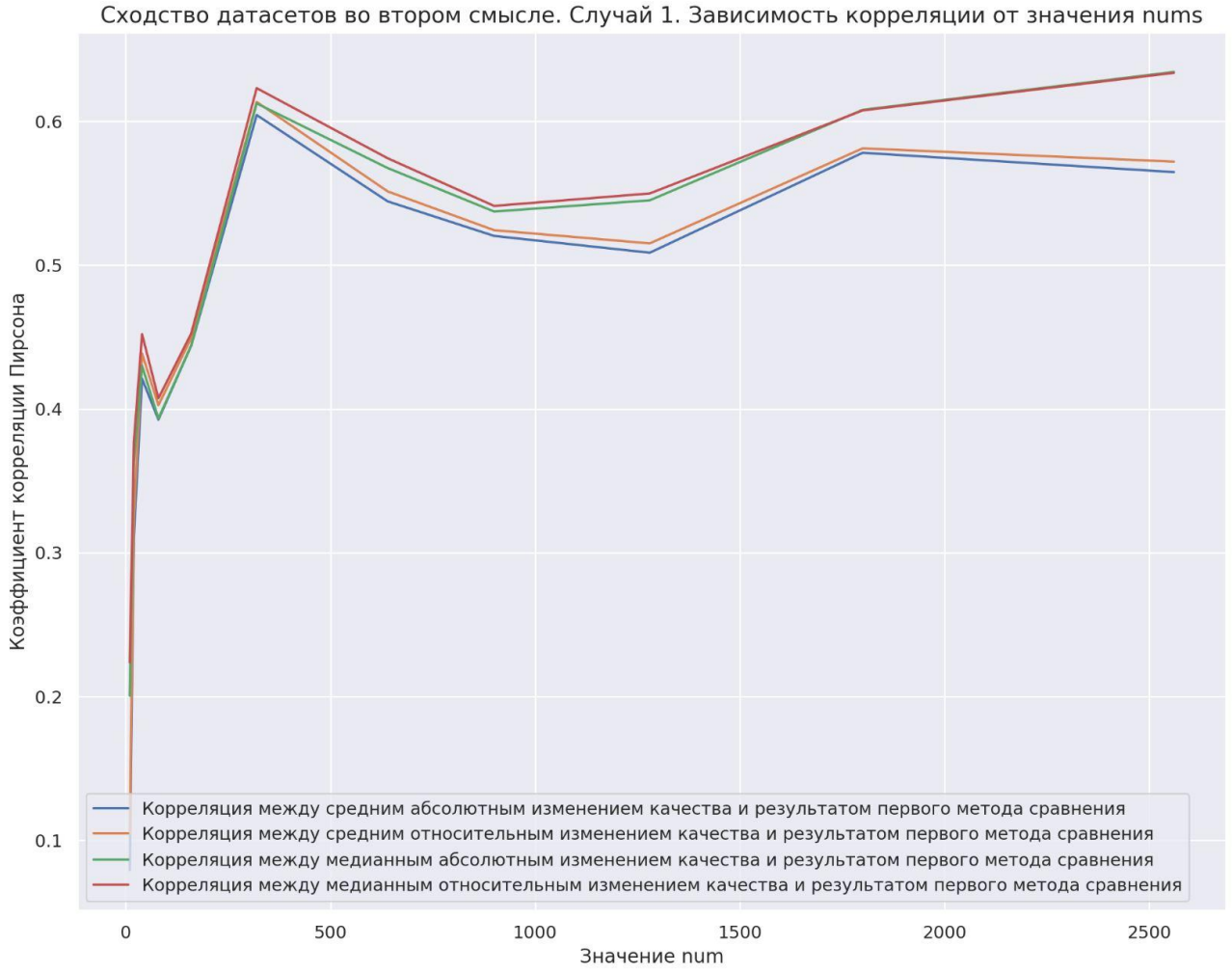
Для случая 2: Коэффициент корреляции, в зависимости от выбранного значения num и того, какое конкретно изменение качества рассматривается, принимает значения от 0.11 до 0.54. Сначала с увеличением значения num значение коэффициента корреляции несколько падает, потом немного колеблется, а затем растет. Оптимальным также является num=2560, при котором значение коэффициента корреляции составляет от 0.41 до 0.54. Таким образом, при оптимальном значении num корреляция получается средняя, ближе к слабой, чем к высокой.

Подводя итог, при оптимальном значении num получается высокая корреляция для случая 1 и средняя корреляция для случая 2. **Метод можно считать эффективным для оценки сходства датасетов в первом смысле.**

Для сходства датасетов во втором смысле:

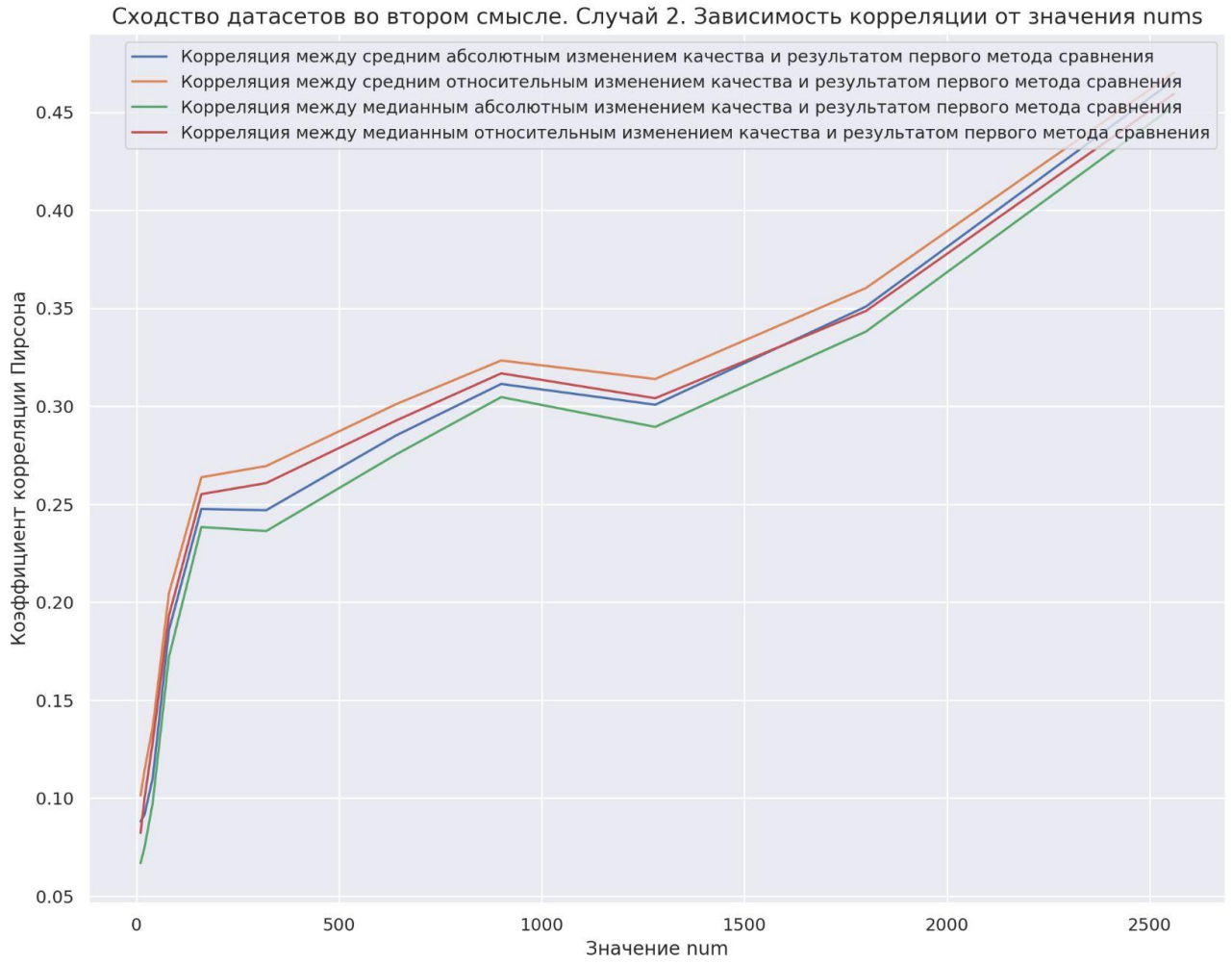
Случай 1: рассматриваются не все возможные пары датасетов, а половина из них: есть пара, соответствующая первой серии экспериментов, есть пара, соответствующая второй серии экспериментов, также есть пары, соответствующие переходам с датасета первой серии экспериментов на датасет второй серии экспериментов. Среди рассматриваемых пар выполнено, что если есть пара (A, B), то пары (B, A) нет среди рассматриваемых.

Рисунок 15



Случай 2: рассматриваются все возможные пары различных датасетов. Иными словами, добавлен переход в обратную сторону.

Рисунок 16



Вывод:

Для случаев 1 и 2 изменения качества (среднее и медианное изменение абсолютного качества, среднее и медианное изменение относительного качества) ведут себя по-разному.

Для случая 1: Коэффициент корреляции, в зависимости от выбранного значения num и того, какое конкретно изменение качества рассматривается, принимает значения от 0.07 до 0.63. Сначала с увеличением значения num значение коэффициента корреляции возрастает с незначительными колебаниями, достигая локального максимума при nums=320. При nums=320 каждый коэффициент корреляции превосходит 0.6, то есть наблюдается стойкая средняя корреляция. С дальнейшим ростом значения num коэффициенты корреляции сначала снижаются, затем повышаются, а в конце ведут себя по-разному. При nums = 2560 значения коэффициентов корреляции, соответствующих медианным изменениям качества, получаются чуть лучше, чем при num=320, а значения коэффициентов корреляции, соответствующих средним изменениям качества – несколько ниже.

В целом, наилучшие результаты для сходства датасетов во втором смысле достигаются при nums=320 и при nums=2560, при которых есть стойкая средняя корреляция.

Для случая 2: Коэффициент корреляции, в зависимости от выбранного значения num и того, какое конкретно изменение качества рассматривается, принимает значения от 0.07 до 0.47. С увеличением значения num значение коэффициента корреляции возрастает с незначительными колебаниями. Оптимальным является num=2560, при котором значение коэффициента корреляции составляет от 0.45 до 0.47. Таким образом, при оптимальном значении num корреляция получается чуть ниже среднего.

Подводя итог, при оптимальном значении num получается стойкая средняя корреляция для случая 1, но для случая 2 корреляция получается ниже среднего. **Метод является условно эффективным для оценки сходства датасетов во втором смысле.**

Экспериментальная часть 7 – Результаты исследования эффективности для второго метода

Для сходства датасетов в первом смысле:

Случай 1: рассматриваются все возможные пары датасетов, в том числе пары с совпадающими датасетами.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.719517
correlation_for_median_absolute_quality_difference_and_second_method_result	0.700215
correlation_for_mean_relative_quality_difference_and_second_method_result	0.702173
correlation_for_median_relative_quality_difference_and_second_method_result	0.689387

Случай 2: рассматриваются все возможные пары *различных* датасетов.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.137801
correlation_for_median_absolute_quality_difference_and_second_method_result	0.118253
correlation_for_mean_relative_quality_difference_and_second_method_result	0.072547
correlation_for_median_relative_quality_difference_and_second_method_result	0.082060

Вывод:

Для случая 1: Коэффициенты корреляции получаются около 0.7, что находится на границе средней и высокой корреляции.

Для случая 2: Коэффициенты корреляции получаются существенно меньше: от 0.7 до 0.14, что соответствует очень слабой корреляции.

Подводя итог, получается стойкая средняя корреляция для случая 1, но для случая 2 корреляция получается очень низкая. **Метод является слабо эффективным для оценки сходства датасетов в первом смысле.**

Для сходства датасетов во втором смысле:

Случай 1: рассматриваются не все возможные пары датасетов, а половина из них: есть пара, соответствующая первой серии экспериментов, есть пара, соответствующая второй серии экспериментов, также есть пары, соответствующие переходам с датасета первой серии экспериментов на датасет второй серии экспериментов. Среди рассматриваемых пар выполнено, что если есть пара (A, B), то пары (B, A) нет среди рассматриваемых.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	0.039861
correlation_for_median_absolute_quality_difference_and_second_method_result	0.121839
correlation_for_mean_relative_quality_difference_and_second_method_result	0.070966
correlation_for_median_relative_quality_difference_and_second_method_result	0.149076

Случай 2: рассматриваются все возможные пары различных датасетов. Иными словами, к случаю 1 добавлен переход в обратную сторону.

	value
correlation_for_mean_absolute_quality_difference_and_second_method_result	-1.519744e-17
correlation_for_median_absolute_quality_difference_and_second_method_result	2.017388e-17
correlation_for_mean_relative_quality_difference_and_second_method_result	2.364859e-03
correlation_for_median_relative_quality_difference_and_second_method_result	1.037911e-02

Вывод:

Для случая 1: Коэффициенты корреляции получаются от 0.04 до 0.15, что соответствует очень слабой корреляции.

Для случая 2: Коэффициенты корреляции получаются менее 0.015, это говорит об отсутствии корреляции.

Подводя итог, получается очень слабая корреляция для случая 1, для случая 2 корреляция получается чрезвычайно близкой к нулю. **Метод является**

абсолютно неэффективным для оценки сходства датасетов во втором смысле.

Замечание: На самом деле, не удивительно, что этот метод совершенно не подходит для оценки сходства датасетов во втором смысле. Легко заметить, что метод симметричный: если методу в качестве аргументов передать на вход датасеты в другом порядке, то возвращаемое методом численное значение не изменится. С другой стороны, сходство датасетов во втором смысле симметричным не является. Датасеты в случае 2 специально подобраны таким образом, чтобы подтвердить или опровергнуть значимость описанного выше несоответствия. Видим, что получилось подтвердить важность описанного выше несоответствия.

Экспериментальная часть 8 – Краткие выводы о выдвинутых гипотезах

Гипотеза 1 (про схожесть топов наиболее часто встречающихся слов) подтвердилась. Разработанный в соответствии с ней метод является эффективным для сходства датасетов в первом смысле, условно эффективным для сходства датасетов во втором смысле.

Гипотеза 2 (про схожесть усредненных эмбеддингов) не подтвердилась. Разработанный в соответствии с ней метод является слабо эффективным для сходства датасетов в первом смысле, абсолютно неэффективным для сходства датасетов во втором смысле.

Заключение

Для каждой из N = 10 рассматриваемых языковых моделей проведено две серии экспериментов (в серии экспериментов два датасета, которые решают близкие задачи бинарной классификации) по описанной в практической части методике. Результаты экспериментов собраны и визуализированы. Сделаны наблюдения, выводы. Сформулированы две гипотезы относительно того, какие признаки могут быть связаны со сходством датасетов (в первом или втором смысле понимания сходства датасетов).

В соответствии с выдвинутыми гипотезами реализованы методы оценки сходства датасетов. Каждый метод оценки сходства датасетов представляет собой функцию, которая принимает два датасета и возвращает численное значение. Возвращаемое численное значение тем больше, чем лучше происходит переход с первого датасета на второй.

Для реализованных методов по заранее сформулированным критериям проверена эффективность – насколько сильно возвращаемые методами численные значения коррелируют со средним и медианным изменением качества при переходе с одного датасета на другой. По результатам проверки подтверждена первая гипотеза и опровергнута вторая.

Лучший из разработанных методов оценки сходства датасетов получился *эффективным* для оценки сходства датасетов в первом смысле и *условно эффективным* для оценки сходства во втором смысле. Данный метод опирается на гипотезу номер 1 (про схожесть наиболее часто встречающихся слов).

Таким образом, поставленные задачи выполнены в полном объеме, цель работы достигнута. Весь код и другие относящиеся к работе материалы выложены в гитхаб-репозиторий [9].

Перспективы дальнейшего исследования:

1) исследовать эффективность хорошо зарекомендовавшего себя метода *на других задачах*, отличных от бинарной классификации: например, на задаче многоклассовой классификации.

   Интересно было бы выяснить, останется ли метод эффективным при переходе к другому классу задач. Исследование можно проводить по той же методике, по которой оно проводилось для задач бинарной классификации.

2) провести дополнительное исследование для неподтвердившейся гипотезы номер 2.

   Как я ранее объяснил в “Экспериментальной части”, разработанный в соответствии с гипотезой номер 2 метод является симметричным: если

методу в качестве аргументов передать на вход датасеты в другом порядке, то возвращаемое методом численное значение не изменится. С другой стороны, сходство датасетов в первом смысле не обязательно является симметричным, а сходство датасетов во втором смысле – точно не является симметричным в общем случае.

Разработанный метод является симметричным потому, что в нём усредненные векторы сравниваются с помощью cosine similarity. Если же сравнивать методы неким *несимметричным* образом, то результаты могут стать совершенно другими. Это, в свою очередь, может повысить эффективность метода, построенного в соответствии со второй гипотезой.

Список источников

[1]. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // arxiv.org. 2018. URL: https://arxiv.org/abs/1810.04805. DOI: 10.48550/ARXIV.1810.04805.

[2]. Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. // arxiv.org. 2018. URL: https://arxiv.org/abs/1804.07461. DOI: 10.48550/ARXIV.1804.07461.

[3]. Rajpurkar, Pranav and Zhang, Jian and Lopyrev, Konstantin and Liang, Percy. SQuAD: 100,000+ Questions for Machine Comprehension of Text. // arxiv.org. 2016. URL: https://arxiv.org/abs/1606.05250. DOI: 10.48550/ARXIV.1606.05250.

[4]. Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. RoBERTa: A Robustly Optimized BERT Pretraining Approach. // arxiv.org. 2019. URL: https://arxiv.org/abs/1907.11692. DOI: 10.48550/ARXIV.1907.11692.

[5]. Lai, Guokun and Xie, Qizhe and Liu, Hanxiao and Yang, Yiming and Hovy, Eduard. RACE: Large-scale ReAding Comprehension Dataset From Examinations. // arxiv.org. 2017. URL: https://arxiv.org/abs/1704.04683. DOI: 10.48550/ARXIV.1704.04683.

[6]. Clark, Kevin and Luong, Minh-Thang and Le, Quoc V. and Manning, Christopher D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. // arxiv.org. 2020. URL: https://arxiv.org/abs/2003.10555. DOI: 10.48550/ARXIV.2003.10555.

[7]. Feng, Fangxiaoyu and Yang, Yinfei and Cer, Daniel and Arivazhagan, Naveen and Wang, Wei. Language-agnostic BERT Sentence Embedding. // arxiv.org. 2020. URL: https://arxiv.org/abs/2007.01852. DOI: 10.48550/ARXIV.2007.01852.

[8]. Ссылка на TensorFlow Hub: https://tfhub.dev/.

[9]. Ссылка на посвященный курсовой работе гитхаб репозиторий: https://github.com/Oleg13Kopylov/comparing_datasets.