Проект 3. Решить задачу DaNetQA / BoolQ

Выполнил: Леушкин Антон Дмитриевич

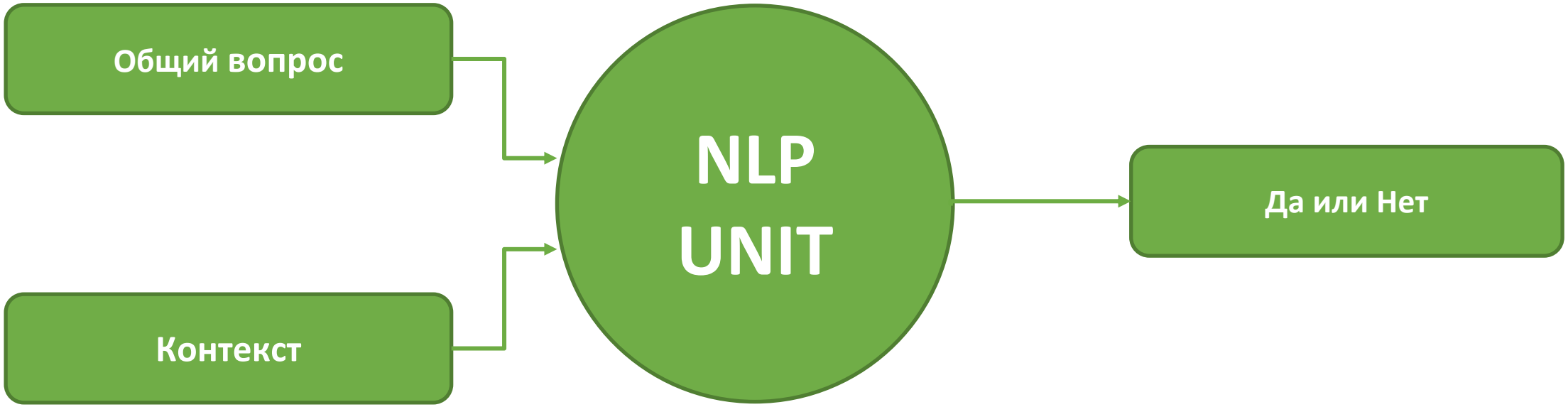
Цель:
Провести анализ задачи DaNetQA - boolQ для русскоязычного набора данных.
Проанализировать существующие решения и попытаться воспроизвести результат.

Задачи в рамках сессий:
- изучить набор данных;
- изучить существующие решения;
- обработать данные;
- воспроизвести две модели и обучить их на подготовленном наборе данных;
- сравнить полученные результаты.

Класс задач NLP

- Распознавание речи из аудио
- Генерация текста и речи
- Анализ текста
- Подсказки при наборе
- Обработка текста
- Извлечение информации из текста
- Автоматическое обобщение или пересказ текста
- Машинный перевод.

DaNetQA – задача NLP бинарной классификации



Постановка задачи DaNetQA: есть вопрос, на который можно ответить да или нет (общий вопрос) и текст, относящийся к нему (контекст). Требуется ответить на вопрос Да или Нет.

Данные

«Вы́ставочный центр» — станция Московского монорельса. Расположена между станциями «Улица Академика Королёва» и «Улица Сергея Эйзенштейна». Находится на территории Останкинского района Северо-Восточного административного округа города Москвы. Переход на станцию ВДНХ Калужско-Рижской линии. Названа в честь Всероссийского выставочного центра — названия ВДНХ с 1992 по 2014 год. 20 ноября 2004 года линия монорельса начала работать в «экскурсионном режиме» и перевезла первых пассажиров .

Вднх - это выставочный центр?

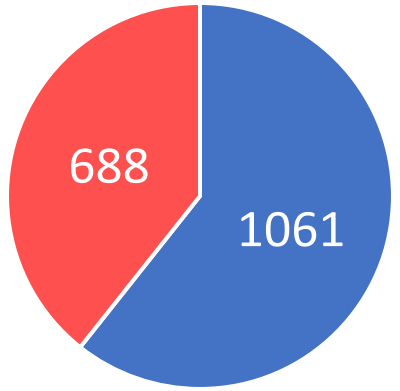
Да

Отметив некоторые недостатки и в целом удачную конструкцию, специалисты ГАУ не рекомендовали принимать ПП Калашникова на вооружение по технологическим причинам. Заключение гласило: С 1942 года Калашников работал на Центральном научно-исследовательском полигоне стрелкового и миномётного вооружения ГАУ РККА. Здесь в 1944 году он создал опытный образец самозарядного карабина, который, хотя и не вышел в серийное производство, частично послужил прототипом для создания автомата. С 1945 года Михаил Калашников начал разработку автоматического оружия под промежуточный патрон 7,62×39 образца 1943 года. Автомат Калашникова победил в конкурсе 1947 года и был принят на вооружение.

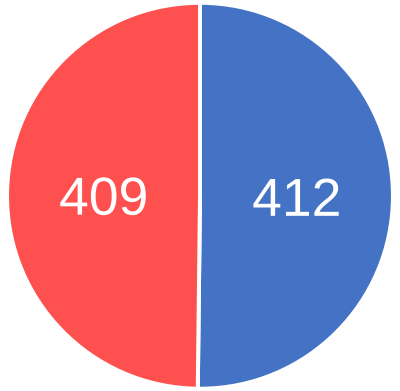
Был ли автомат калашникова в вов?

Нет

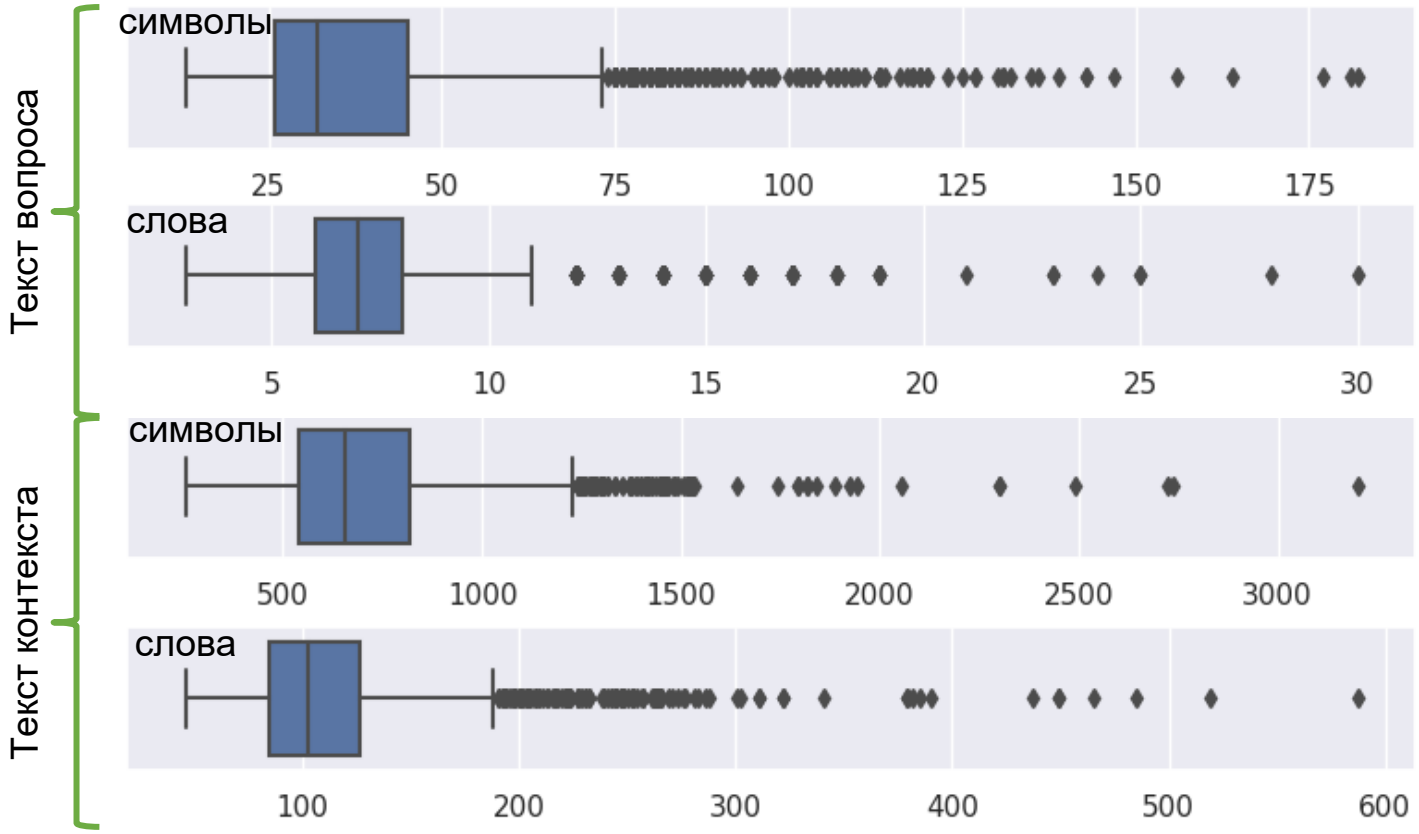
Train Dataset



Test Dataset



Анализ данных



Текст вопроса	количество символов	количество слов
MIN	14	3
MEAN	39	7
MAX	182	30

Текст контекста	количество символов	количество слов
MIN	254	46
MEAN	708	113
MAX	3200	587

Критическая масса кортежей состоит из вопроса в 25-50 символов или 7 слов и контекста в 550-700 символов или 70-130 слов. Но есть выбросы, доходящие до 182 символов или 30 слов в вопросе и 3200 символов и 587 слов.

Очистка данных

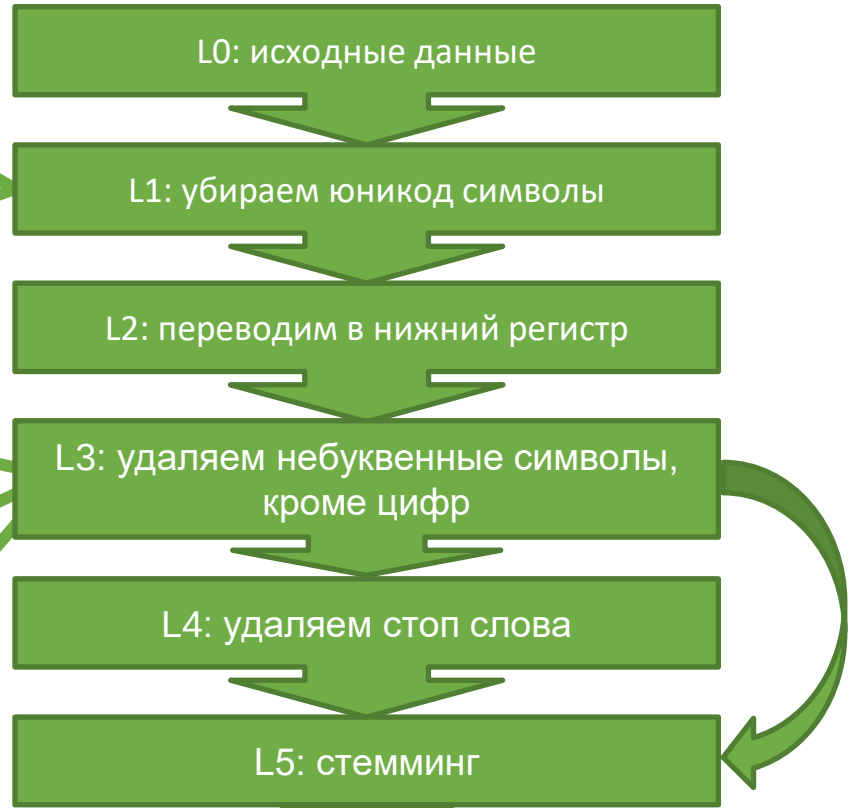
symbol	˘	¨	´	`	.	‘	^	‑
count	27745	5295	1618	6	6	5	4	4

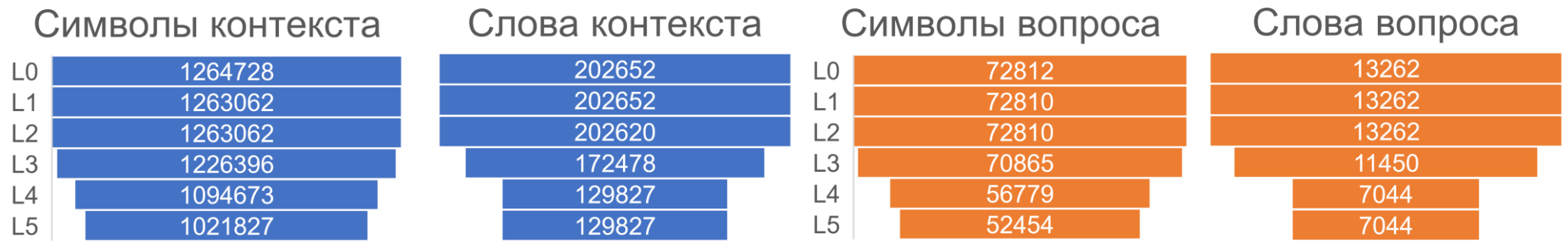
symbol	‑	’	°	̊	~	ˇ	̥	˜
count	4	3	2	2	2	1	1	1

symbol	,	.	—	-	?	«	»	)
count	27875	19367	4524	4010	3439	2246	2233	1430

symbol	:	(	%	/	]	[	!	>
count	1395	1337	348	198	154	149	54	19

symbol	*	$	=	|	<	#	&	{	}
count	13	10	9	7	6	6	5	3	1



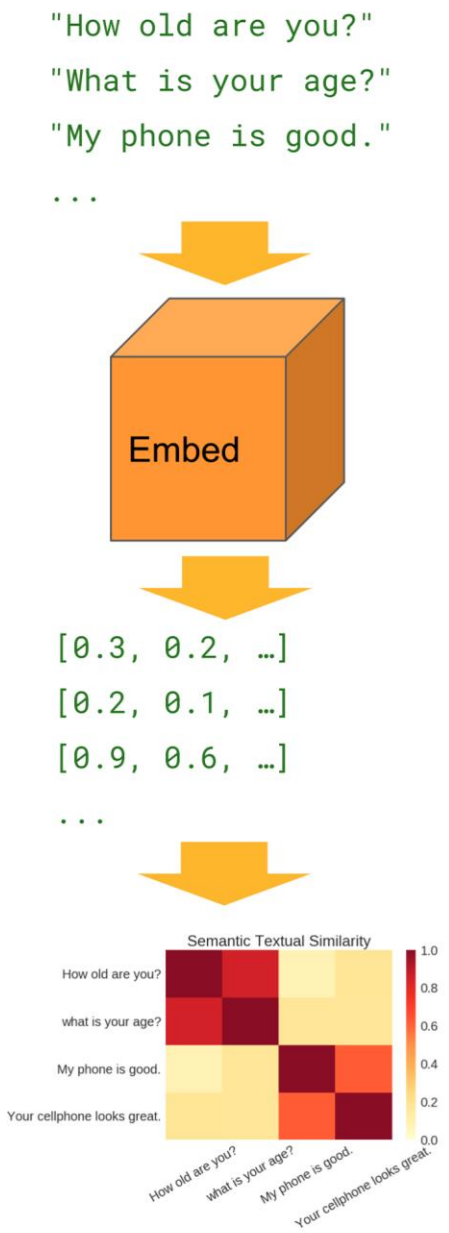


Пример процесса очистки данных

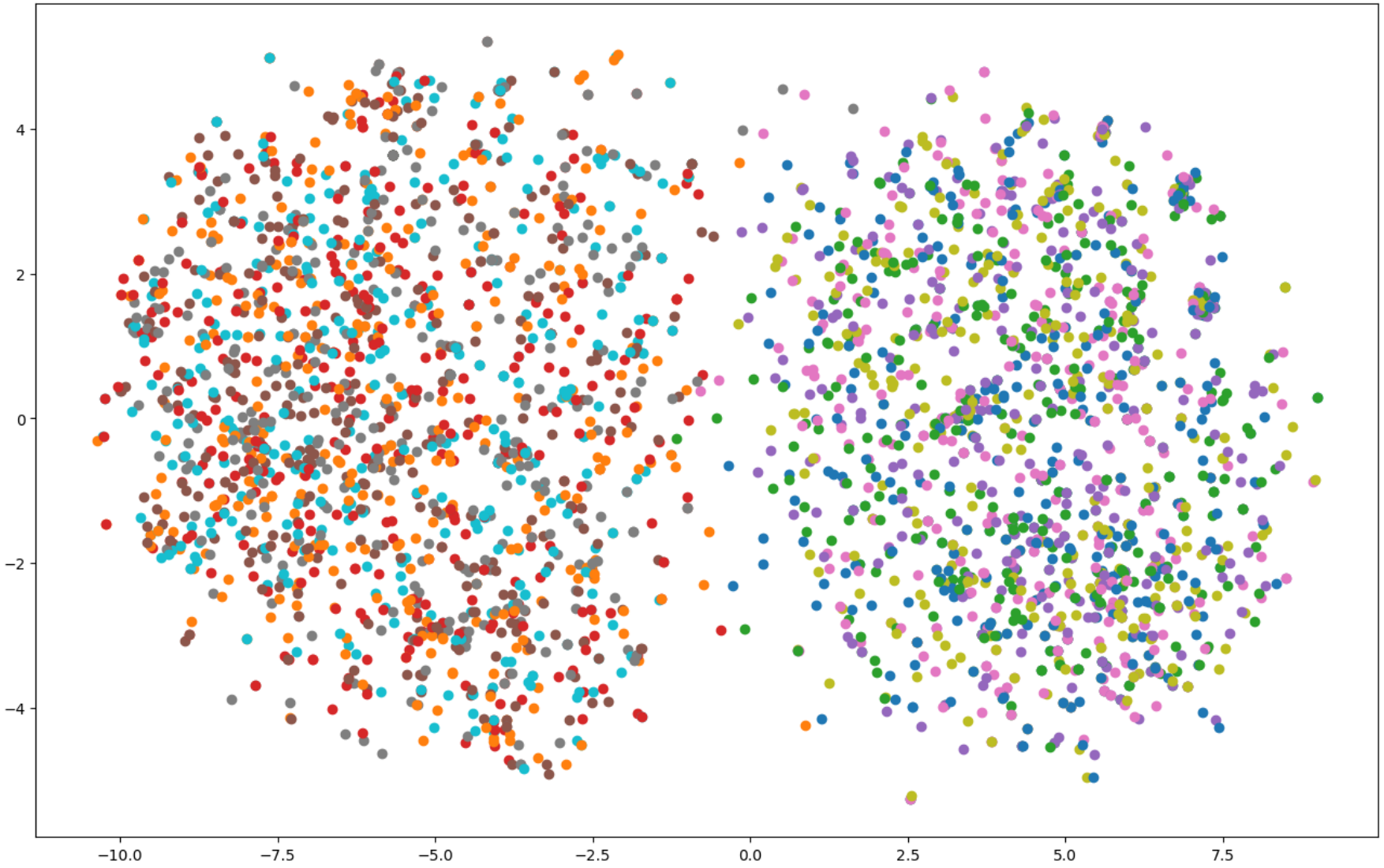
Уровень очистки	Пример контекста	Пример вопроса
L0: исходные данные	«Вы́ставочный центр» — станция Московского монорельса. Расположена между станциями «Улица Академика Королёва» и «Улица Сергея Эйзенштейна». Находится на территории Останкинского района Северо-Восточного административного округа города Москвы. Переход на станцию ВДНХ Калужско-Рижской линии. Названа в честь Всероссийского выставочного центра — названия ВДНХ с 1992 по 2014 год. 20 ноября 2004 года линия монорельса начала работать в «экскурсионном режиме» и перевезла первых пассажиров .	Вднх - это выставочный центр?
L1: убираем юникод символы	«Выставочныи центр» — станция Московского монорельса. Расположена между станциями «Улица Академика Королева» и «Улица Сергея Эизенштеина». Находится на территории Останкинского раиона Северо-Восточного административного округа города Москвы. Переход на станцию ВДНХ Калужско-Рижскои линии. Названа в честь Всероссииского выставочного центра — названия ВДНХ с 1992 по 2014 год. 20 ноября 2004 года линия монорельса начала работать в «экскурсионном режиме» и перевезла первых пассажиров .	Вднх - это выставочныи центр?
L2: переводим в нижний регистр	«выставочныи центр» — станция московского монорельса. расположена между станциями «улица академика королева» и «улица сергея эизенштеина». находится на территории останкинского раиона северо-восточного административного округа города москвы. переход на станцию вднх калужско-рижскои линии. названа в честь всероссииского выставочного центра — названия вднх с 1992 по 2014 год. 20 ноября 2004 года линия монорельса начала работать в «экскурсионном режиме» и перевезла первых пассажиров .	вднх - это выставочныи центр?
L3: удаляем служебные символы и пунктуацию	выставочныи центр станция московского монорельса расположена между станциями улица академика королева и улица сергея эизенштеина находится на территории останкинского раиона северо восточного административного округа города москвы переход на станцию вднх калужско рижскои линии названа в честь всероссииского выставочного центра названия вднх с 1992 по 2014 год 20 ноября 2004 года линия монорельса начала работать в экскурсионном режиме и перевезла первых пассажиров	вднх это выставочныи центр
L4: удаляем стоп слова	выставочныи центр станция московского монорельса расположена станциями улица академика королева улица сергея эизенштеина находится территории останкинского раиона северо восточного административного округа города москвы переход станцию вднх калужско рижскои линии названа честь всероссииского выставочного центра названия вднх 1992 2014 год 20 ноября 2004 года линия монорельса начала работать экскурсионном режиме перевезла первых пассажиров	вднх это выставочныи центр
L5: стемминг	выставочны центр станц московск монорельс располож станц улиц академик королев улиц серге эизенштеин наход территор останкинск раион север восточн административн округ город москв переход танц вднх калужск рижско лин назва чест всероссииск выставочн центр назван вднх 19922014 год 20 ноябр 2004 год лин монорельс нача работа экскурсион режим перевезл перв пассажир	вднх эт выставочны центр

Семантический анализ данных

Мультиязычная модель «universal-sentence-encoder-multilingual» делает кодировку для целого предложения, в отличии от word2vec FastText и TF-IDF, что позволяет проверить корреляцию между контектом и общим вопросом.



T-SNE Диаграмма



Кортежи с низкой корреляцией

L5 кортеж с корреляцией -0.06

в современно росс встреча случа когд женщин воспитыва ребенк без отц дают ем вмест отчеств матрон эт практик не призна законодательн однак загс идут навстреч так пожелан голомидов марин васильевн русск антропонимическ систем на рубеж век вопрос ономастик главны редактор а к матве ответственны секретар л а феоктистов екатеринбург издательств уральск университет 2005

разреш ли в росс матчеств

Да

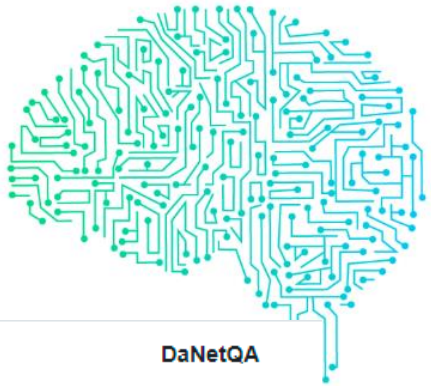
L4 кортеж с корреляцией -0.06

билли сближается своеи приемнои семьеи появляется школе шазам поддержать фредди тому приводит вместе собои самого супермена сцене время титров сивана своеи тюремнои камере одержимости рисует стене загадочные символы прерывает мистер маинд ранее находившиися заключении скалы вечности рассказывает столпах зла которые соидутся вместе будут править семью сферами сцене титров фредди проверяет билли говорить рыбои ссылаясь аквамена билли считает такую способность глупои ашер энджел билли бэтсон подросток которыи превращаться взрослого супергероя произнеся магическое слово шазам оригинале являющееся акронимом шести легендарных богов героев древнего мира способностеи мудрости соломона силы геракла стоикости атланта мощи зевса смелости ахиллеса скорости меркурия дэвид колсмит исполнил роль маленького билли

хищных птицах сцена титров

Да

Существующие решения



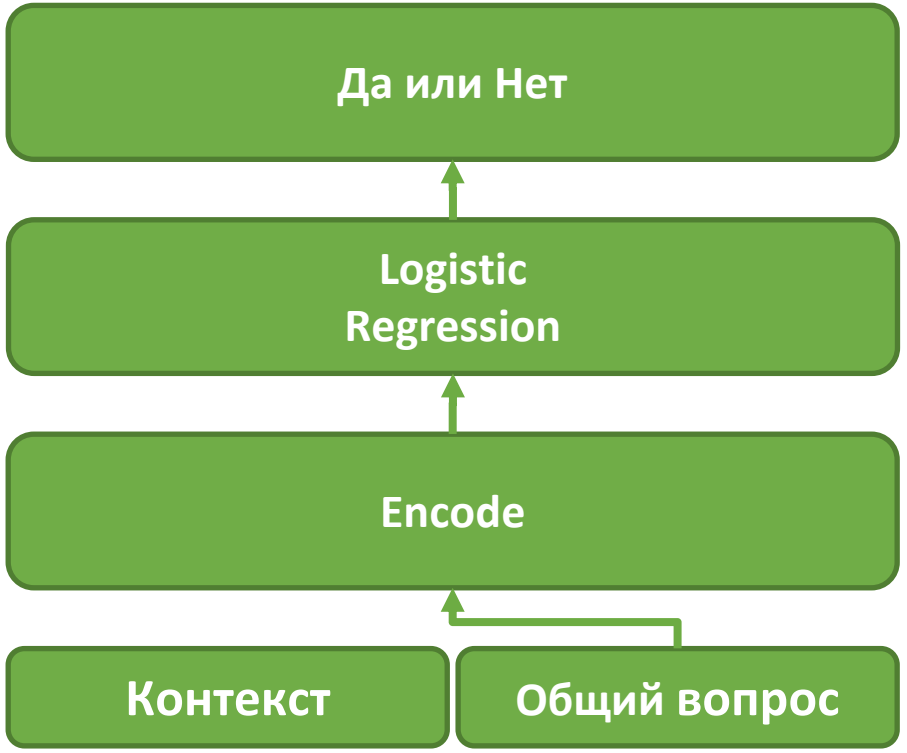
В интернете не так много исследований с решением конкретно задачи DaNetQA, чаще она фигурирет как «одна из» в качестве проверки адаптируемости модели под задачи.

Rank	Name	Team	DaNetQA	Rank	Name	Team	DaNetQA
1	HUMAN BENCHMARK	AGI NLP	0.915	12	RuGPT3XL few-shot	SberDevices	0.59
2	Golden Transformer v2.0	Avengers Ensemble	0.911	13	RuBERT plain	DeepPavlov	0.639
3	YaLM p-tune (3.3B frozen + 40k trainable params)	Yandex	0.85	14	SBERT_Large_mt_ru_finetuning	SberDevices	0.697
4	RuLeanALBERT	Yandex Research	0.76	15	SBERT_Large	SberDevices	0.675
5	ruT5-large finetune	SberDevices	0.79	16	RuGPT3Large	SberDevices	0.604
6	ruRoberta-large finetune	SberDevices	0.82	17	RuBERT conversational	DeepPavlov	0.606
7	Golden Transformer v1.0	Avengers Ensemble	0.917	18	Multilingual Bert	DeepPavlov	0.624
8	ruT5-base finetune	Sberdevices	0.769	19	heuristic majority	hse_ling	0.642
9	ruBert-large finetune	SberDevices	0.773	20	RuGPT3Medium	SberDevices	0.634
10	ruBert-base finetune	SberDevices	0.712	21	RuGPT3Small	SberDevices	0.61
11	YaLM 1.0B few-shot	Yandex	0.637	22	Baseline TF-IDF1.1	AGI NLP	0.621

Таблица результатов таких исследований представлена на russiansuperglue.com [6], актуальна на 09.11.2022

Encoder + LogisticRegression для решения задачи DaNetQA

Уровень очистки данных	TD-IDF Pre-Trained	Universal-Sentence-Encoder-Multilingual
L0: исходные данные	0.59	0.58
L1: убираем юникод символы	0.59	0.58
L2: переводим в нижний регистр	0.59	0.57
L3: удаляем служебные символы и пунктуацию	0.59	0.58
L4: удаляем стоп слова	0.55	0.56
L5: стемминг	0.54	0.57



Лучшая точность доходит до 0.59, это на 0.09 выше чем при использовании псевдослучайных чисел.

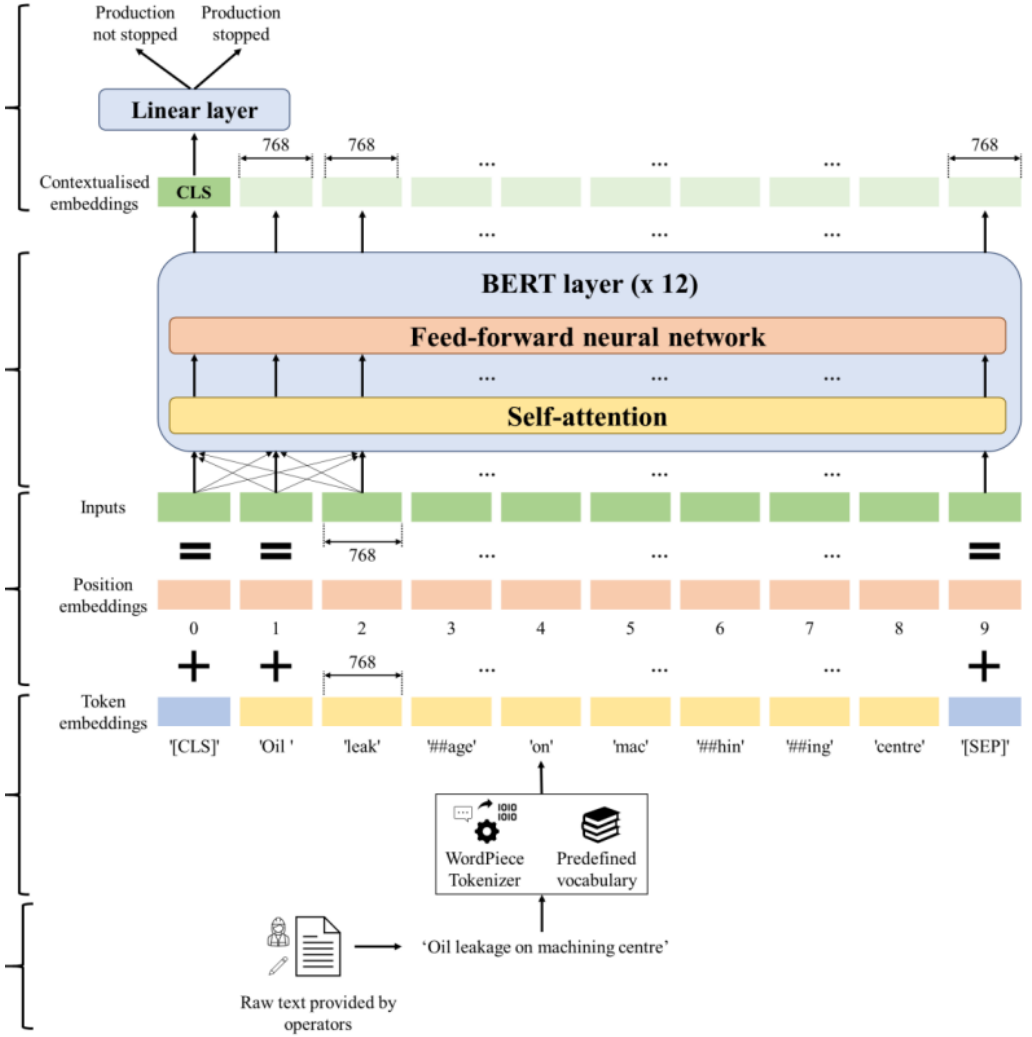
Fine-Tune (Ru)BERT для решения задачи DaNetQA

RuBert-base-cased-sentence [4]

Sentence RuBERT (Russian, cased, 12-layer, 768-hidden, 12-heads, 180M parameters) is a representation-based sentence encoder for Russian.

It is initialized with RuBERT and fine-tuned on SNLI google-translated to russian and on russian part of XNLI dev set.

Sentence representations are mean pooled token embeddings in the same manner as in Sentence-BERT.



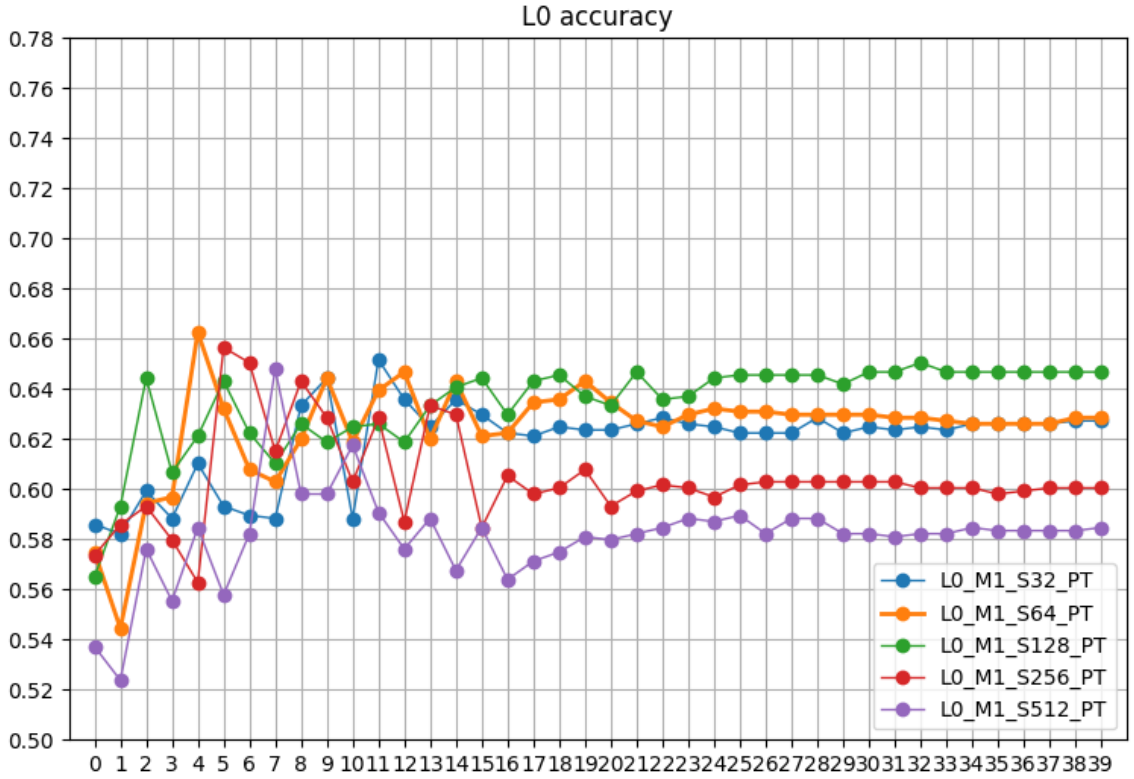
Конфигурация обучения модели

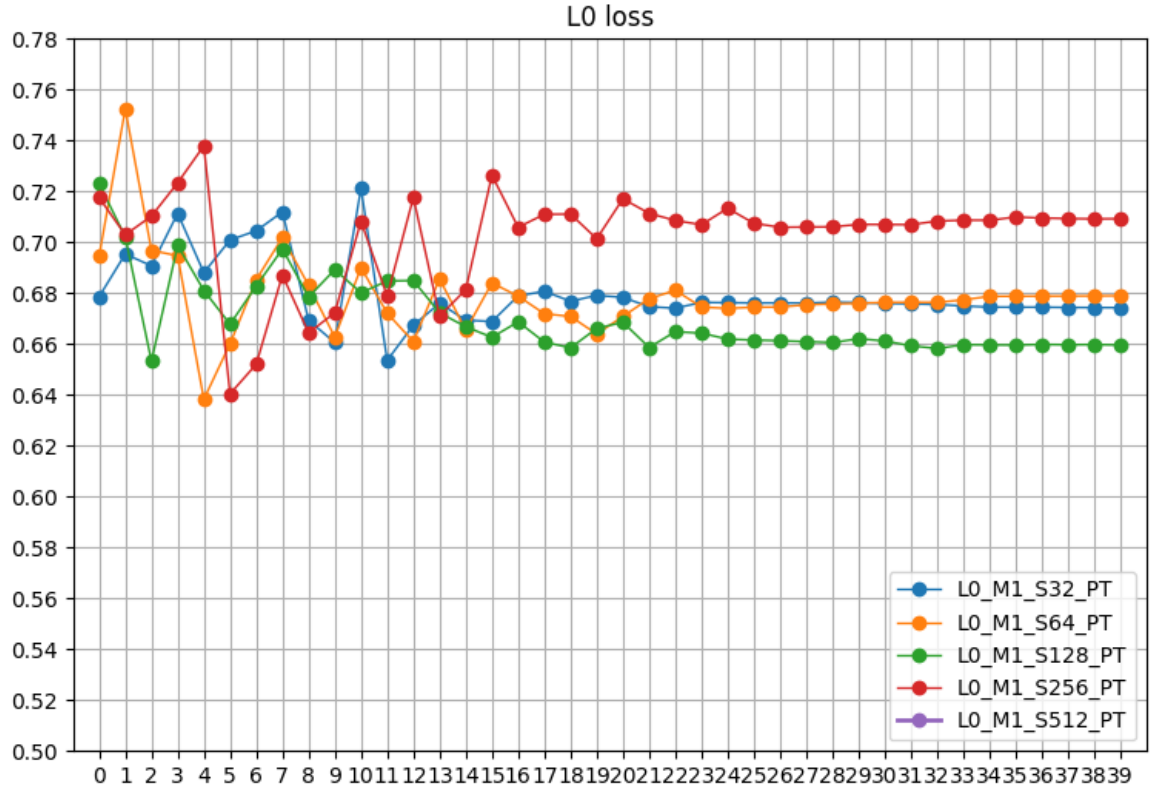
Параметр	Значения	Комментарий
Clean Level	L0, L1, L2, L3, L4, L5	Будем пробовать обучать данные с разным уровнем очистки
Learning Rate	3e-5	Начальное значение скорости обучения
Epochs	40	Типичное кол-во эпох для fine-tune это 2-6, но в исследовании ВШЭ и Сбера [5] используется 40 эпох
Seq Length	32, 64, 128, 256, 512	Длина кодировки, при этом вопрос кодируется полностью, а длина обрезается за счет контекста
Batch Size	16 для Seq Length = 512, 8 для остальных	Влияет в основном на количество требуемой VRAM
Seed	17	Фиксируем сид для воспроизводимости результатов

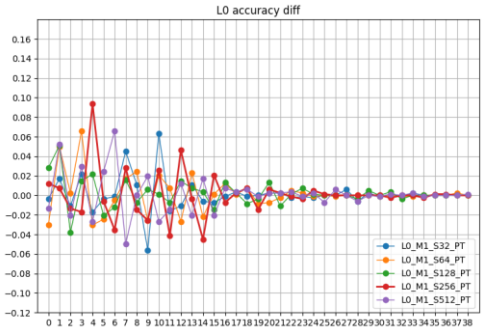
План проведения экспериментов

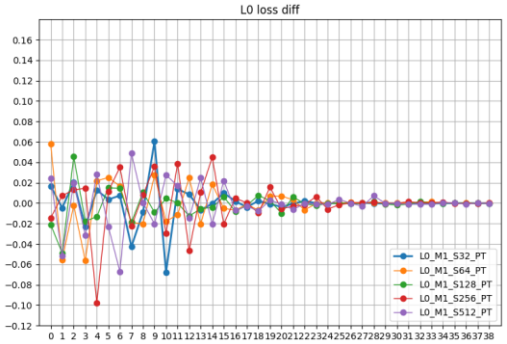
- Провести обучение модели для комбинаций параметров
  - Configs = [*Clean Level x Seq Length*],
  - Получившееся кол-во Configs = 25
- Проанализировать полученные данные о Accuracy и Loss моделей
  - График значений метрики от эпохи
  - График изменения значения метрики между двумя эпохами
- Выбрать лучшую модель
  - Мануально проверить на небольшой выборке валидационных данных
  - Визуализировать confusion matrix для тестовых данных

Clean Level = 0: исходные данные

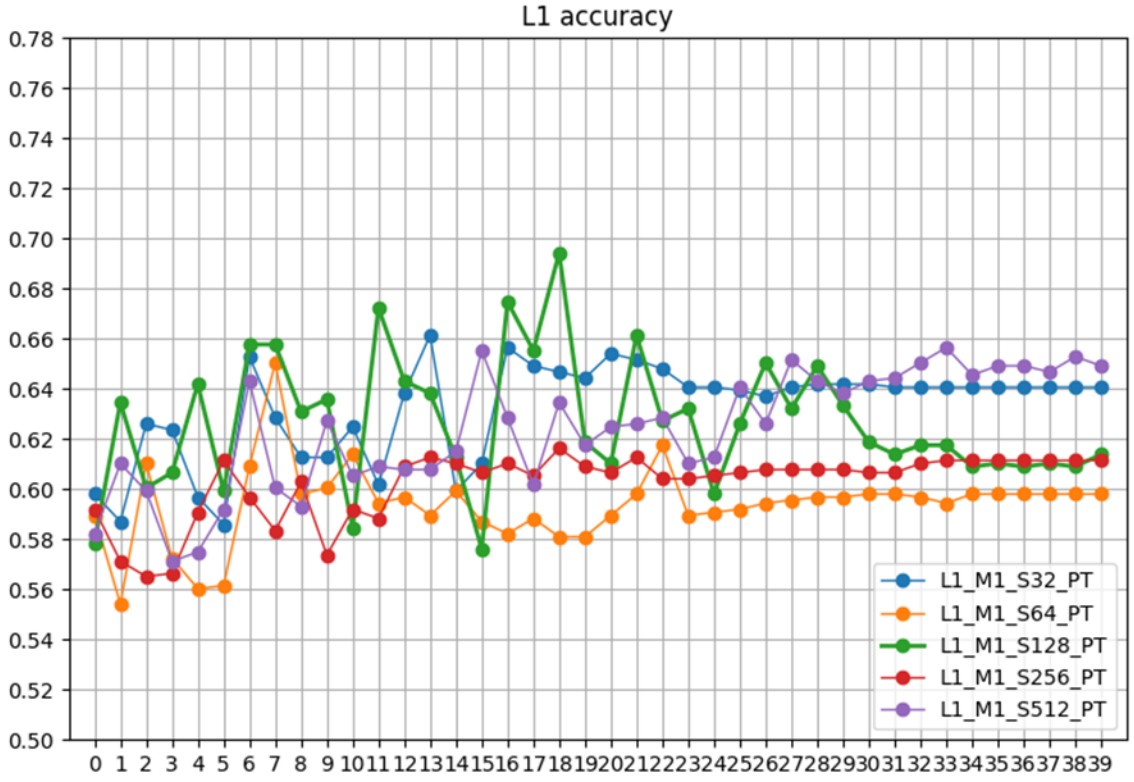


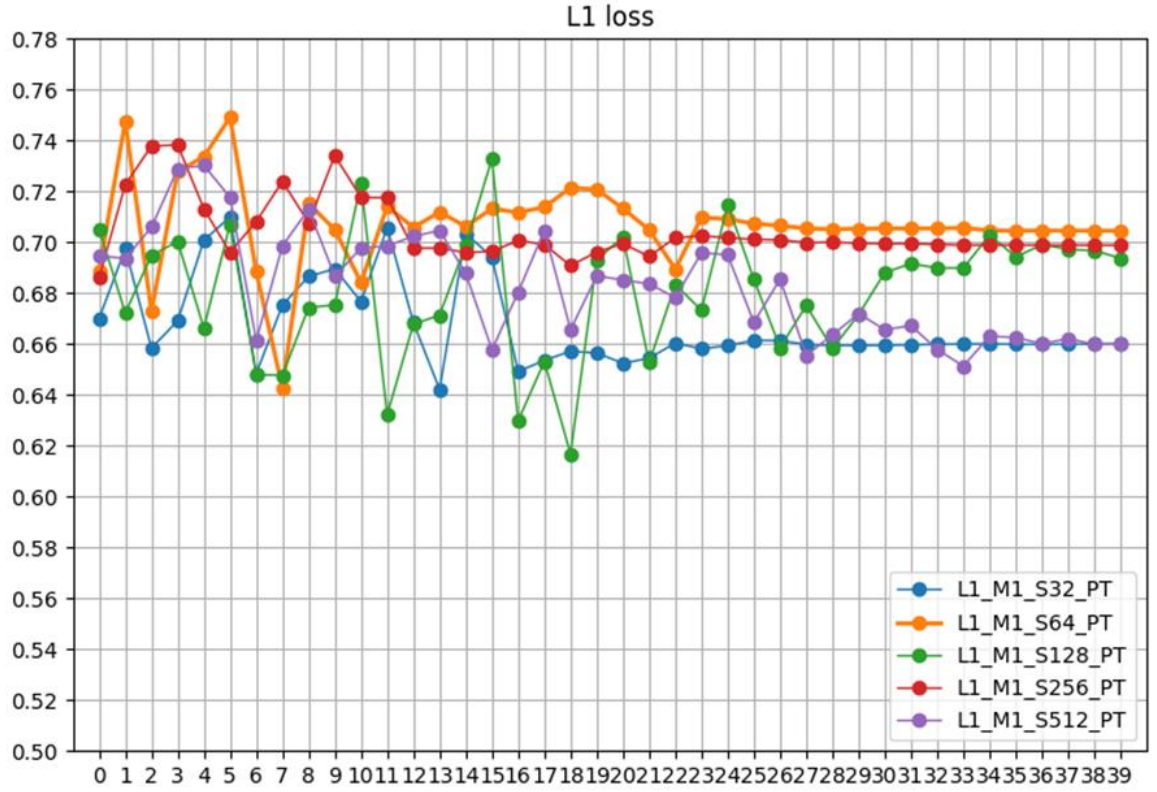


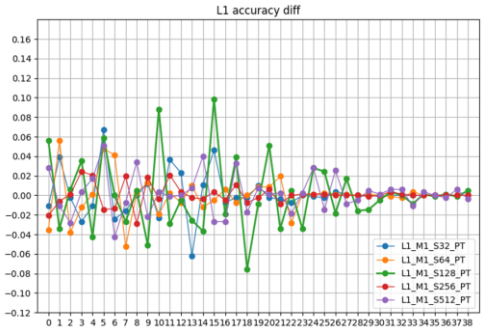


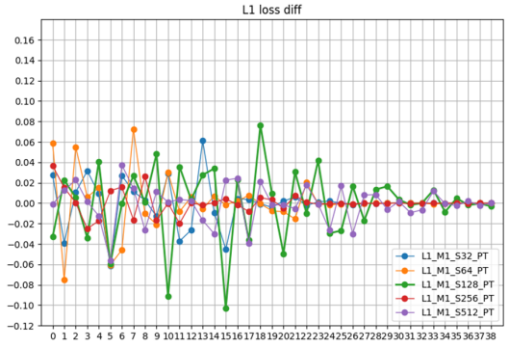


Clean Level = 1: убираем юникод символы

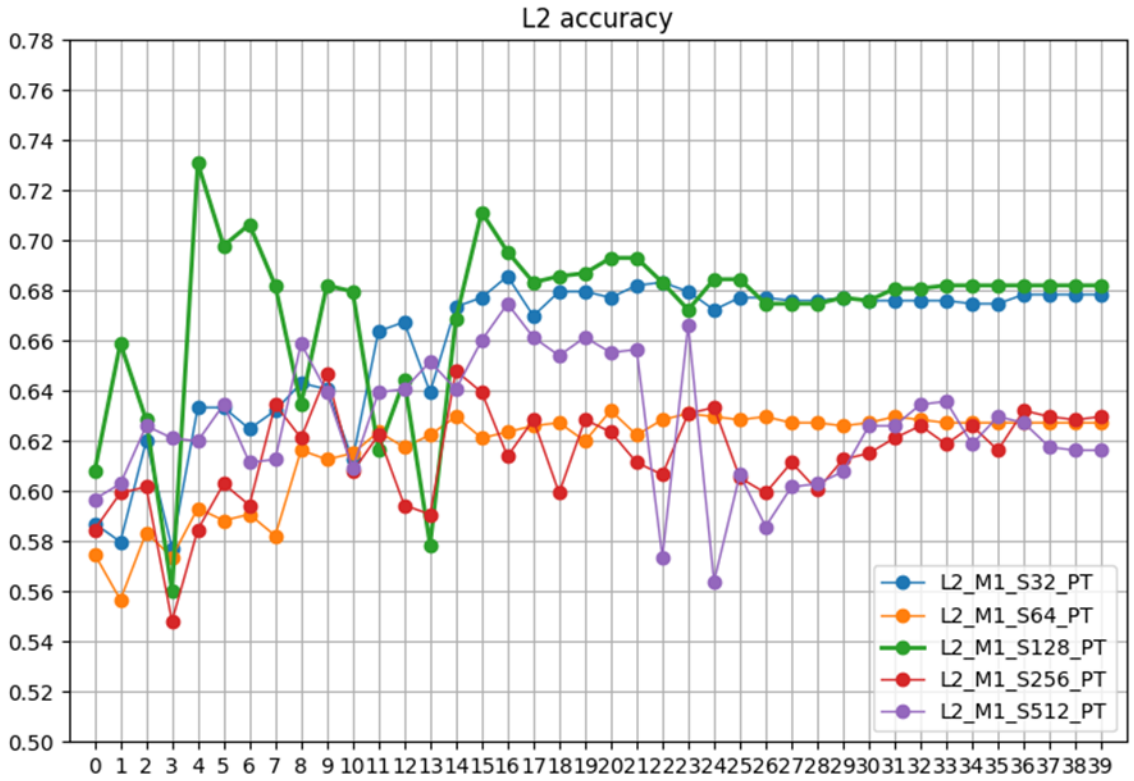


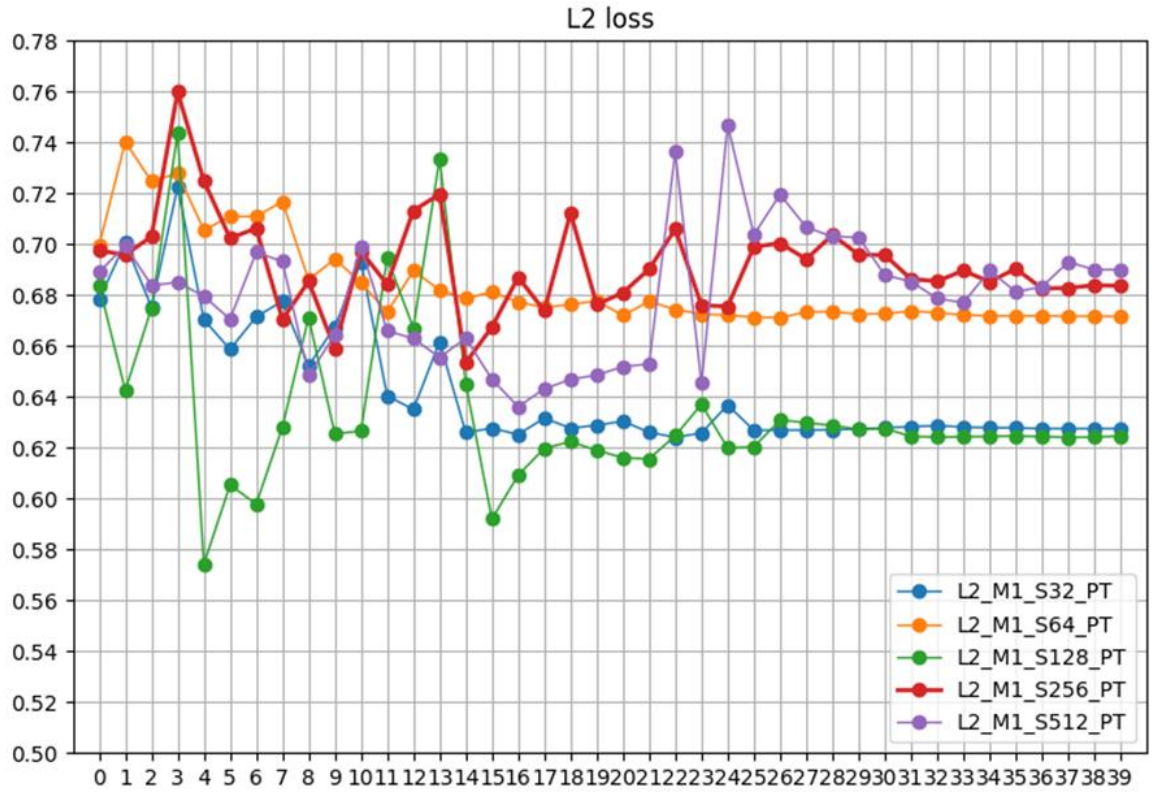


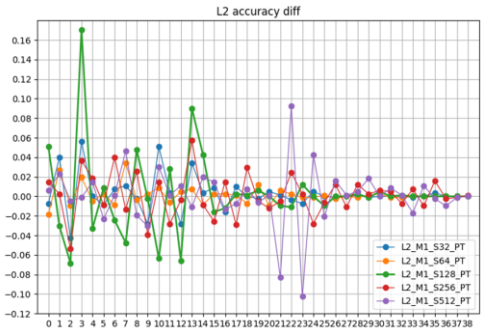


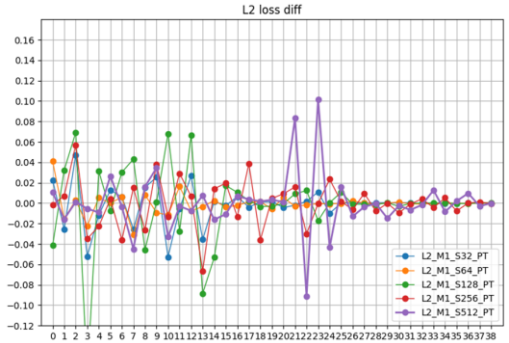


Clean Level = 2: переводим в нижний регистр

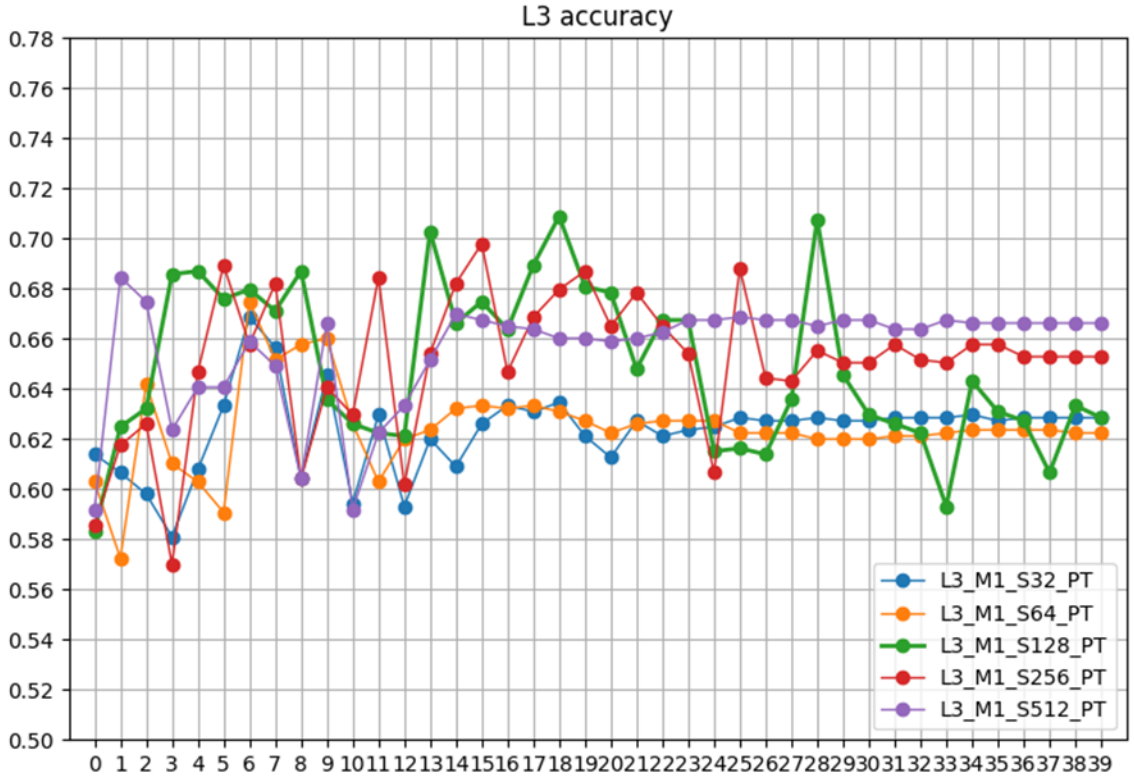


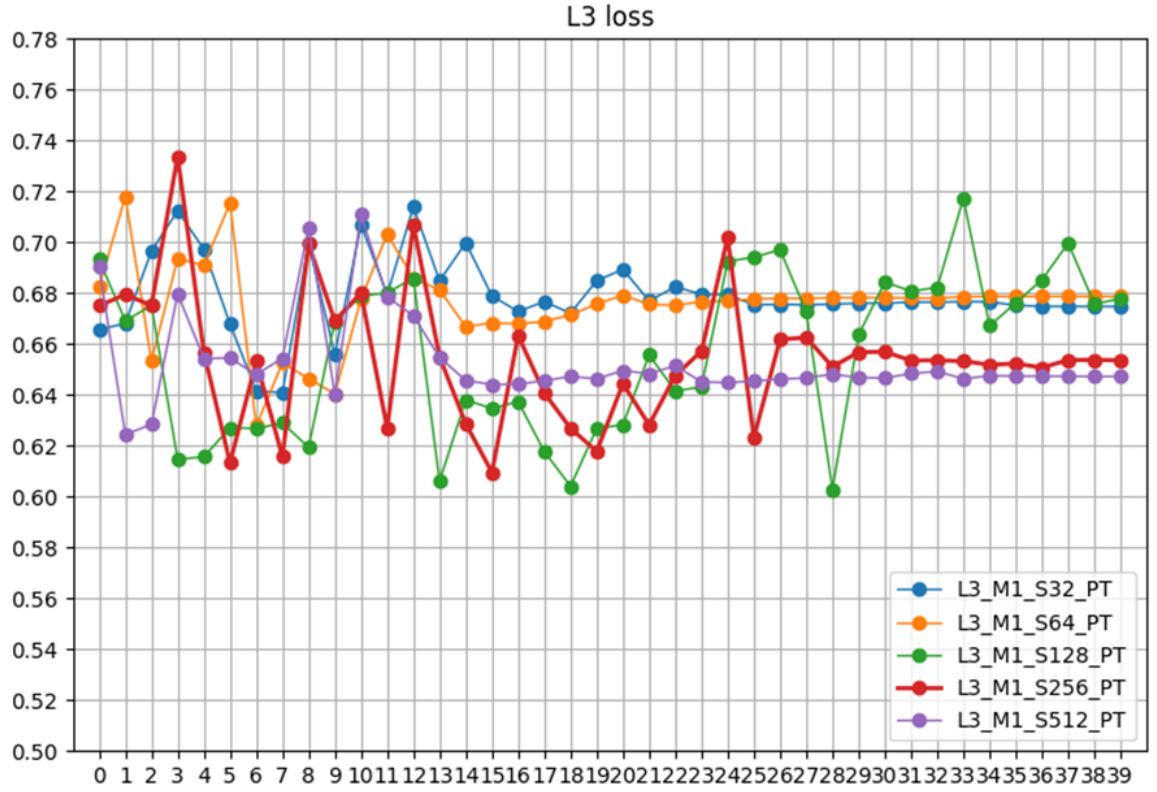


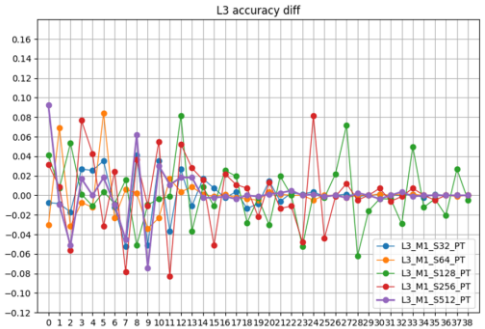


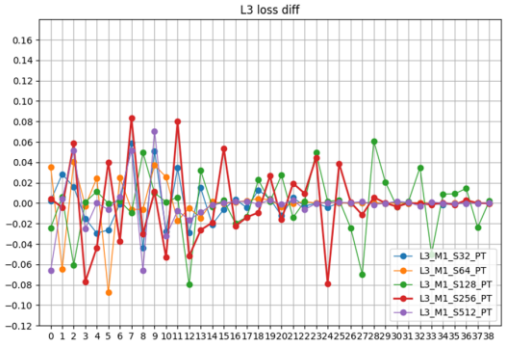


Clean Level = 3: удаляем служебные символы и пунктуацию

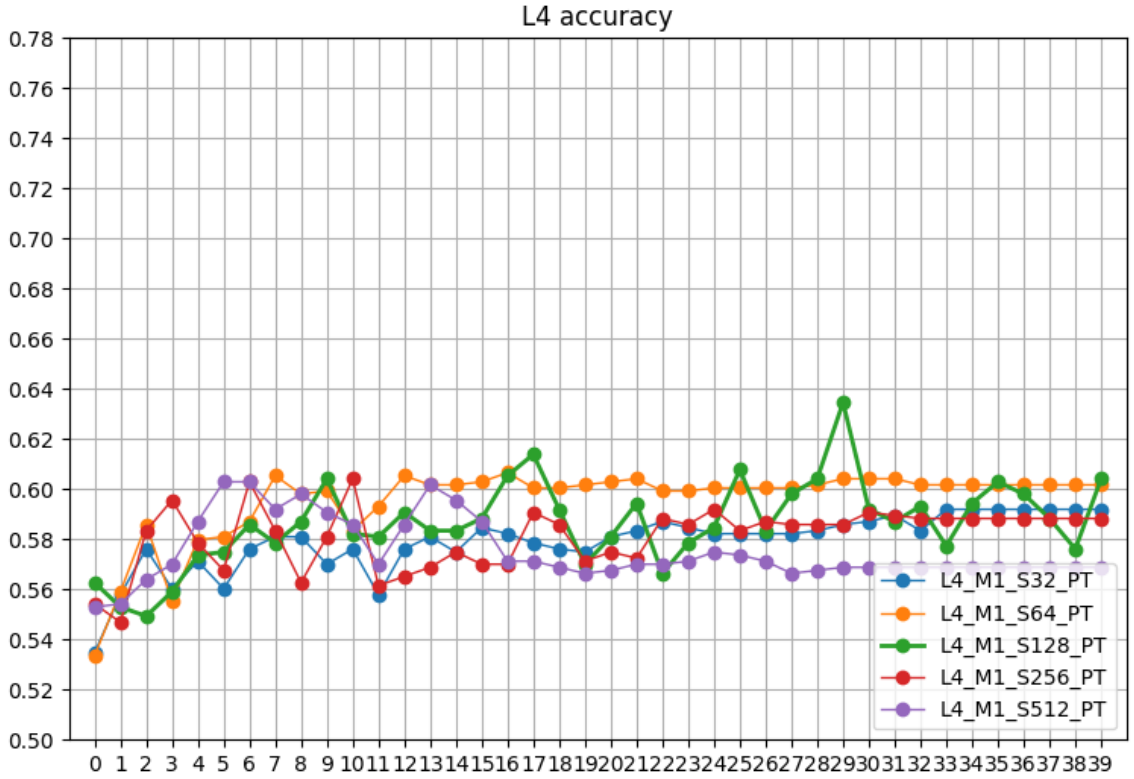


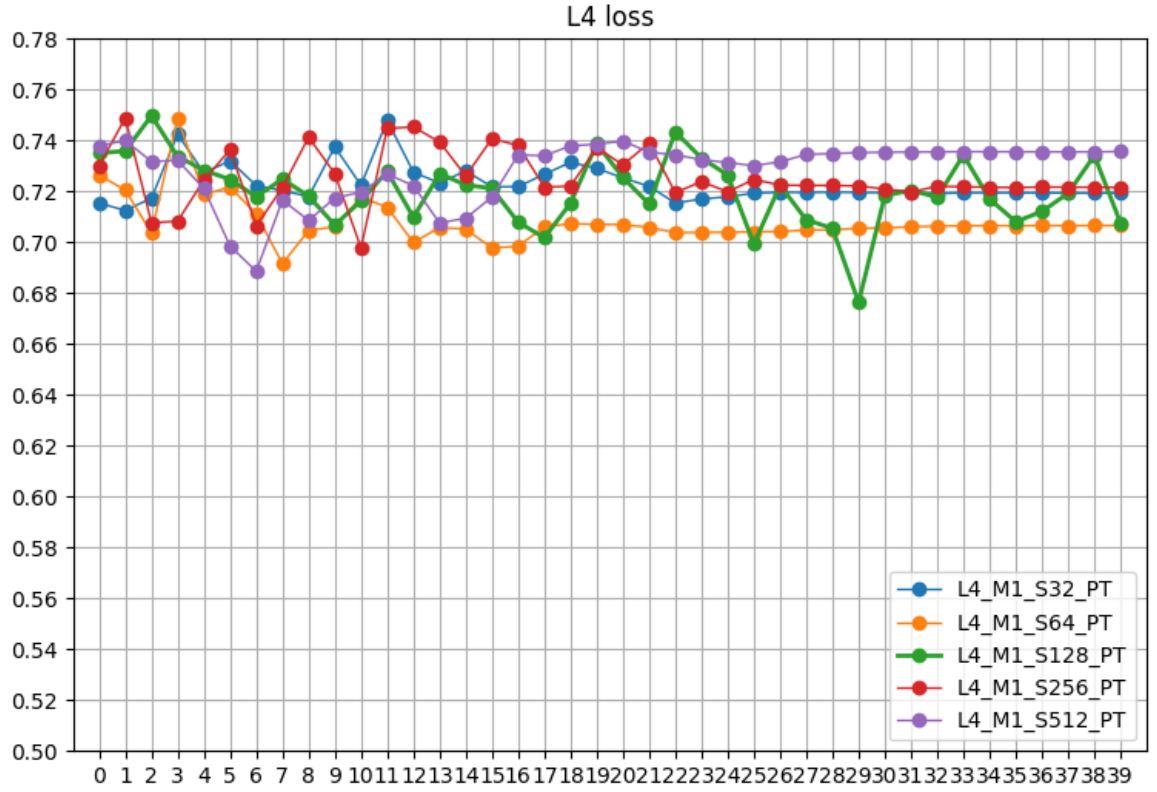


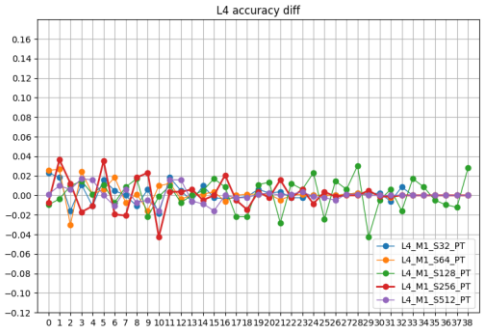


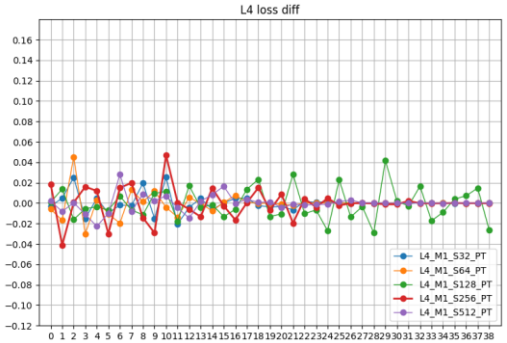


Clean Level = 4: удаляем стоп слова

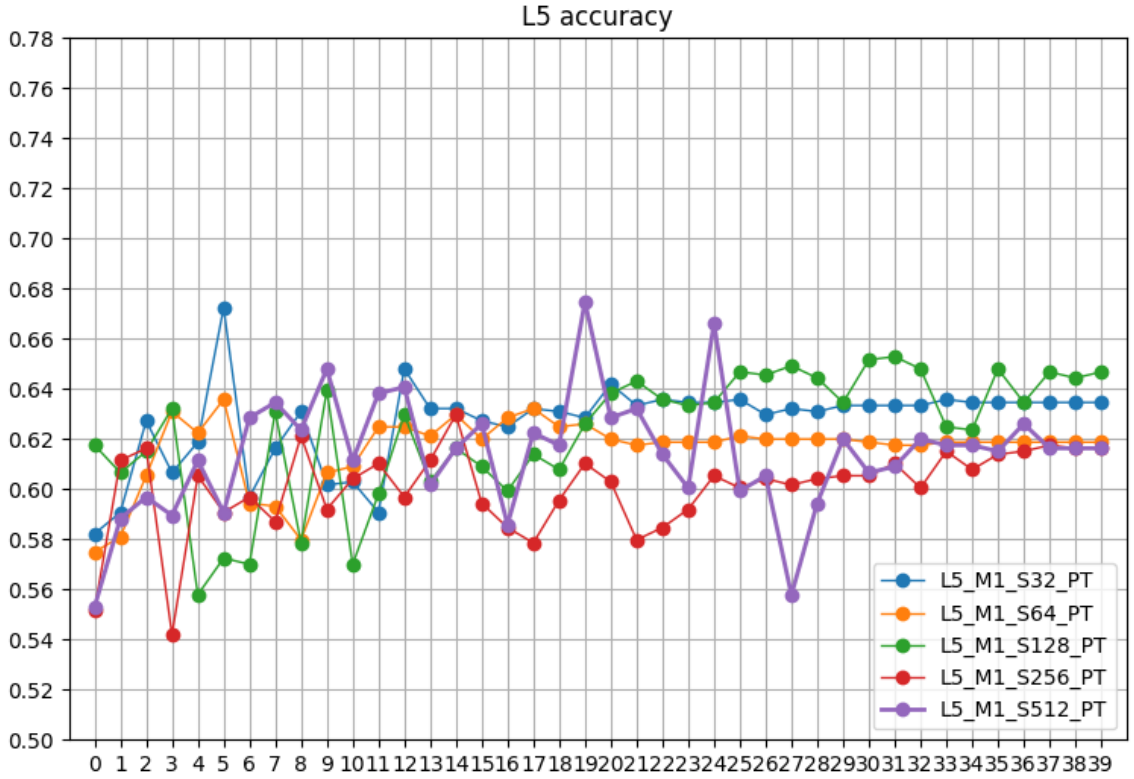


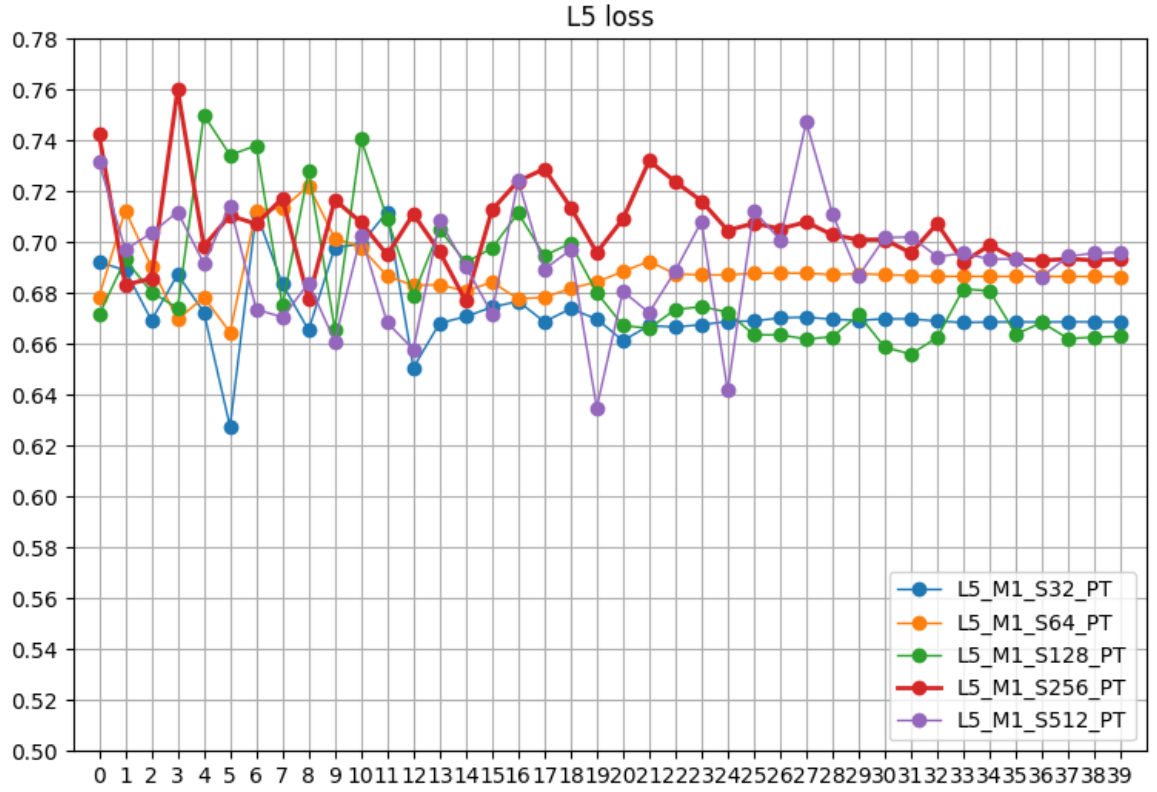


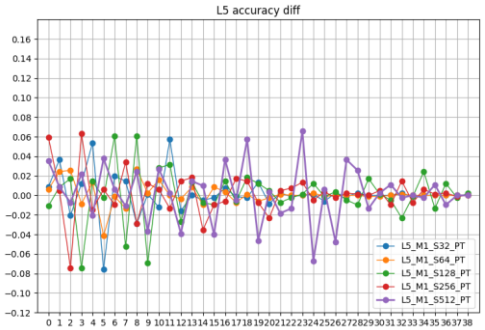


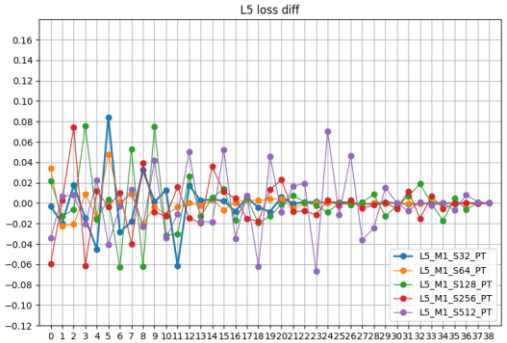


Clean Level = 5: стемминг







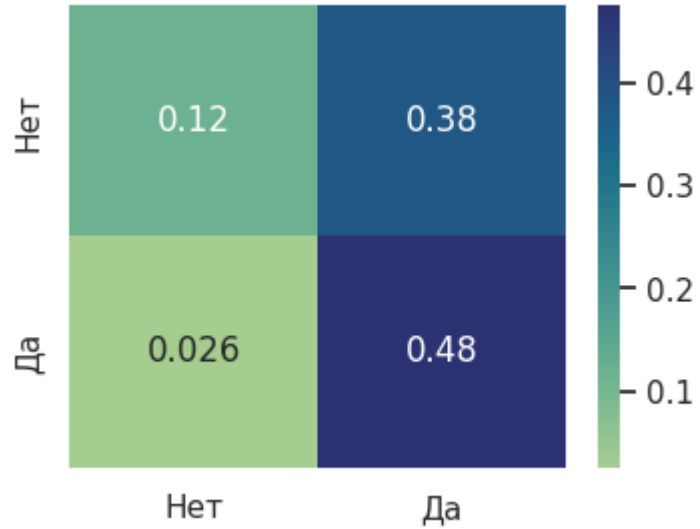


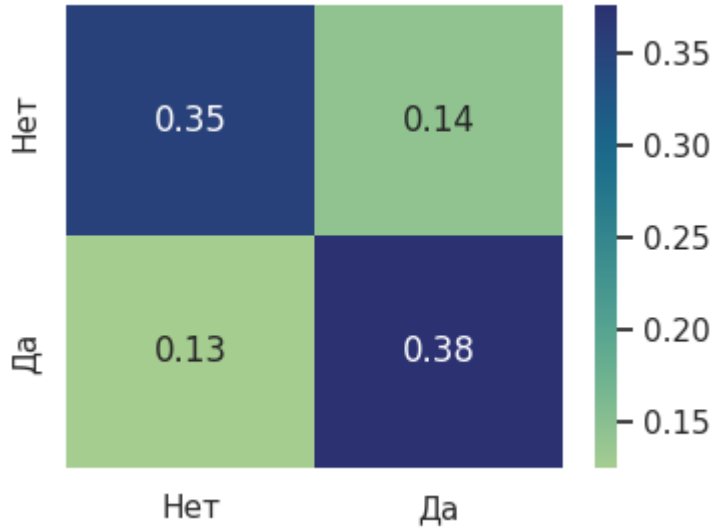
Выбор лучшей модели

С каждой конфигурации мы сохраняли по две модели: последнюю и лучшую по выбранной метрике (Accuracy).
Проверив все была выбрана лучшая. Accuracy, на 5 эпохе равняется **0.73**.
Сравним метрики моделей:

TF-IDF + LogisticRegression				
	Precision	Recall	F1	Support
Нет	0.82	0.24	0.37	409
Да	0.56	0.95	0.70	412
Точн.			0.59	821
Сред.	0.69	0.59	0.53	821

RuBert-Cased				
	Precision	Recall	F1	Support
Нет	0.74	0.71	0.72	409
Да	0.72	0.75	0.74	412
Точн.			0.73	821
Сред.	0.73	0.73	0.73	821





Мануальная проверка модели. Часть 1 из 3

Вопросы и ответы были взяты из валидационного датасета (из начала и хвоста), он сделан для сабмита[6] не содержит ответов на вопросы поэтому он не участвовал в обучении.

Хюррéм Хасеки́-султáн ; настоящее имя неизвестно. В поздней литературной традиции её имя Алексáндра Гаври́ловна Лисóвская; ок. 1502 или ок. 1505 — 15 или 18 апреля 1558) — наложница, а затем жена османского султана Сулеймана Великолепного, хасеки, мать султана Селима II. Документальные источники и даже сколь-либо надёжные письменные свидетельства, говорящие о жизни Хюррем до поступления в гарем, отсутствуют.

Была ли у джихангира наложница? — Да

Гру́ппа кро́ви — описание индивидуальных антигенных характеристик эритроцитов, определяемое с помощью методов идентификации специфических групп углеводов и белков, включённых в мембраны эритроцитов. У человека открыто несколько систем антигенов в разных группах крови. Группы крови различают как у животных, так и у людей. В мембране эритроцитов человека содержится более 300 различных антигенных детерминант, молекулярное строение которых закодировано соответствующими генными аллелями хромосомных локусов. Количество таких аллелей и локусов в настоящее время точно не установлено.

Правда ли у животных нет групп крови? — Нет

Эта статья о группе белков. О пищевом продукте см. : СейтанКлейкови́на, глюте́н — понятие, объединяющее группу запасающих белков, обнаруженных в семенах злаковых растений, в особенности пшеницы, ржи и ячменя. Термином «клейковина» обозначаются белки фракции проламинов и глютелинов. Клейковина была впервые выделена Якопо Бартоломео Беккари в 1728 году из муки.

Есть ли в хлебе белок? — Да

Мануальная проверка модели. Часть 2 из 3

Отравления ртутью — расстройства здоровья, связанные с избыточным поступлением паров или соединений ртути в организм. Токсические свойства ртути известны с глубокой древности. Соединения ртути — киноварь, каломель и сулема — применялись для разных целей, в том числе и в качестве ядов. С древних времён известна также и металлическая ртуть, хотя её токсичность поначалу сильно недооценивалась. Ртуть и её соединения стали особенно широко применяться в средние века, в частности при производстве золота и серебряных зеркал, а также при изготовлении фетра для шляп, что вызвало поток новых, уже профессиональных отравлений.

Полезна ли ртуть с градусника? — Да

Элизабет Глэдис Миллвина Дин — англичанка, которая была последней из выживших пассажиров «Титаника», затонувшего 15 апреля 1912 года и его самой юной пассажиркой. На момент гибели лайнера ей было два с половиной месяца и, соответственно, никаких воспоминаний о трагедии у неё не было. Миллвина родилась 2 февраля 1912 года в Бранскомбе в семье Бертрама Фрэнка Дина и Жоржетты Эвы Лайт. Супруги вдвоём управляли трактиром в Лондоне. У Миллвины был брат Бертрам Вер Дин

Все ли погибли на титанике? — Нет

Нур-Султа́н — столица Республики Казахстан с 10 декабря 1997 года. Город расположен на севере страны, на берегах реки Ишим, административно разделён на 4 района. Акмолинск получил статус города 26 сентября 1862 года. Городом-миллионером Астана стала в июне 2017 года, когда население составило 1 002 874 жителя. На начало 2020 года население Нур-Султана составляло 1 136 156 человек, что является вторым показателем в Казахстане после Алма-Аты.

Правда ли астану переименовали? — Да

Клубневидно вздутые корни весят до 15 кг, содержат 20—40 % крахмала. Растение широко культивируется в тропических регионах, например, в Африке. В пищу идёт похожий на картофелину корнеплод, который может достигать 8 см в диаметре и 1 м в длину, масса — от 3 до 10 кг. В корнеплодах много крахмала. В сыром виде корнеплоды очень ядовиты и употребляются только варёными или печёными.

Есть ли в батате крахмал? — Да

Мануальная проверка модели. Часть 3 из 3

Болотный крокодил, или магер — пресмыкающееся из семейства настоящих крокодилов, обитающее на территории Индостана и прилегающих стран, таких как Пакистан. Это один из трёх обитающих в Индии крокодилов, наряду с гавиалом и гребнистым крокодилом. Ближайший современный родственник болотного крокодила — сиамский крокодил, вместе с гребнистым крокодилом они образуют обособленную азиатскую кладу. У болотного крокодила грубая голова без гребней или выростов чешуйчатых костей, тяжёлые и широкие челюсти, длина которых превышает ширину у основания в 1,3—2,5 раза. Четыре большие пластины на шее образуют квадрат, с меньшими по размерам пластинами на каждой стороне.

Водятся ли в индии крокодилы?

Да

Остров Сент-Маргерит — крупнейший из Леринских островов, расположен на расстоянии менее километра от города Французской Ривьеры Канны. Длина острова с востока на Запад около 3 км, ширина 900 м. Наиболее известный объект острова — форт-тюрьма, в которой в XVII веке содержался человек в железной маске. Первые упоминания об острове, тогда ещё необитаемом, относятся ко временам Древнего Рима, когда остров носил название Леро. Вероятно современное название острову дали крестоносцы, которые построили на острове часовню святой Маргариты Антиохийской. В XIV столетии, вероятно благодаря сочинению Раймонда Феро, остров стал ассоциироваться с вымышленной святой Маргерит, сестрой Гонората Арелатского, основателя монастыря на близлежащем острове Сент-Онора.

Был ли человек в железной маске?

Да

Фауна лесных почв — совокупность видов животных, для которых лесная почва является средой обитания, часть лесной фауны. Животных, обитающих в почве, в зависимости от размеров особей относят к следующим группам: макрофауна — в основном мелкие млекопитающие, в том числе землеройки, кроты. мезофауна — её представляют дождевые черви, многоножки, мокрицы, насекомые, их личинки. микрофауна — нематоды, энхитреиды, клещи, в основном панцирные клещи, ногохвостки и другие. Нанофауна - это одноклеточные простейшие. В любых лесах среди беспозвоночных почв преобладают сапрофаги, которые питаются лесным опадом, грибницей, гниющей древесиной.

Являются ли сапрофаги хищниками?

Да

Возможные улучшения

- Увеличение количества данных
  - Добавление к существующим данным переведенных вопросов из bool Q
  - Синтез абсолютно новых данных
  - Использование других NLP решений, например в качестве вопросов использовать выжимку из текста
- Очистка данных
  - Подбор выборки стоп-слов и служебных символов, которая мы увеличивала, а не ухудшала результат
  - Использование леммитизации вместо стемминга
- Базовая модель
  - Использование large-вариации BERT может дать большую точность
- Процесс обучения
  - Варьирование размера батча.
  - Использование других метрик при выборе лучшей модели

Список литературы

[1] Обучение с использованием Catalyst - Yorko/bert-finetuning-catalyst: Code for BERT classifier finetuning for multiclass text classification (github.com)

[2] Описание архитектуры трансформер - 1706.03762.pdf (arxiv.org)

[3] Описание архитектуры BERT - (PDF) Using deep learning to value free-form text data for predictive maintenance (researchgate.net)

[4] Предобученная модель BERT - DeepPavlov/rubert-base-cased-sentence · Hugging Face

[5] Статья от ВШЭ и СберAI про датасет - DaNetQA: a yes/no Question Answering Dataset for the Russian Language

[6] Таблица результатов существующих моделей - Russian SuperGLUE

[7] Документация Catalyst - catalyst-team/catalyst: Accelerated deep learning R&D (github.com)

Проект 3. Решить задачу DaNetQA / BoolQ

Спасибо за внимание!

Выполнил: Леушкин Антон Дмитриевич