Создание русского DynaSent корпуса

Степанова Ангелина, Фадеева Полина, Ткач Анна, Ревак Ксения

Аннотация

В этой работе была проведена попытка поизвести адаптацию динамического датасета, предложенного в статье DynaSent: A Dynamic Benchmark for Sentiment Analysis [2] на материал русского языка. Были проделаны 5 этапов: сбор русскоязычного корпуса в 197k предложений, файн-тюнинг RoBERTa, анализ и переразметка вызвавших затруднение у модели отзывов, обучение небольшой модели на переразмеченных данных, анализ возникшей поломки, дообучение.

1 Введение

В оригинальной статье представляется DynaSent - новый динамический бенчмарк для анализа тональности на английском языке, включающий 121k предложений. Идея динамического бенчмарка заключается в том, что поэтапно происходит доразметка и дообучение моделей на новой разметке, конкретно на случаях, вызвавших у модели сложности. Таким образом авторы предлагают алгоритм постоянного усложнения задачи для модели, благодаря чему новые модели смогут лучше обучаться и лучше справляться с задачами сентиментного анализа в реальном применении.

В данной работе была проделана попытка также задать основу динамически развивающегося бэнчмарка, но на материале русского языка. В отличие от авторов оригинальной статьи у нас не было доступа к 1000 аннотаторов, из-за чего мы выбрали другой подход и столкнулись с непредвиденными сложностями, на данный момент происходит продолжение дообучения модели второго уровня.

2 Этап 1. Сбор корпуса

В первую очередь были изучены существующие русскоязычные датасеты для сентиментного анализа и для данной работы отобраны и объединены датасеты RuReviews (90k) с отзывами с маркетплэйсов о женской одежде и аксессуарах [4] и Russian Hotel Reviews (57k) с отзывами об отелях [3]. Так как мы стремились к трехклассовому корпусу (положительный, отрицательный и нейтральный сентимент), как в оригинальной статье, а в датасетах отзывы оценивались звездочками, то было произведено следующее преобразование: 1,2 звезды → отрицательный сентимент, 3 → нейтральный, 4,5 → положительный. Получившееся распределение лэйблов в изначальном датасете можно увидеть на Рис. 1.

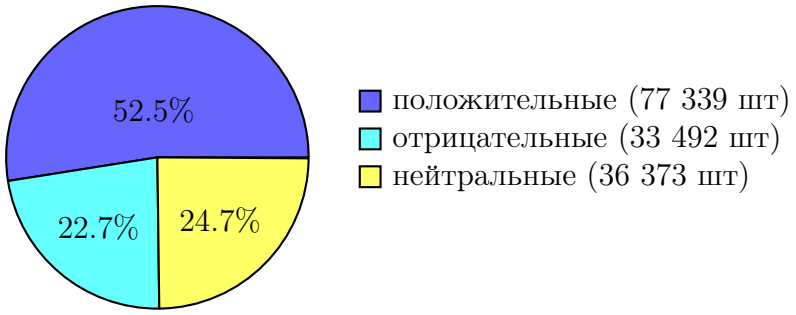

Рис. 1: Распределение лэйблов в собранном на первом этапе датасете.

3 Этап 2. Обучение модели и результаты

Для файн-тьюнинга была выбрана модель RoBERTa[1], потому что она демонстрирует высокую эффективность в задачах обработки естественного языка. [1] Она была дообучена на собранном корпусе, параметры обучения были выставлены такие же, как у модели в оригинальной статье.

Датасет был разделен на обучающий и тестовый в соотношении 0.8 к 0.2, таким образом в тестовом датасете оказалось 29 410 отзывов. Для измерения качества полученной модели была использована F1 мера, как стандартная метрика, позволяющая при датасете с неравномерно распределенными классами получить достаточно полное представление о качестве классификации. В Таблице. 1 представлены результаты по лэйблам, а также F1 мера по каждому из классов. По всем классам результат F1 = 0.849, что значительно выше, чем F1 = 0.756 у первой модели на первом датасете в оригинальной статье.

true label / predicted label	отрицательный	нейтральный	положительный
отрицательный	**5085**	1217	93
нейтральный	*1458*	**5405**	841
положительный	81	751	**14480**
F1	0.781	0.717	0.943

Таблица 1: Распределение примеров в тестовом датасете по настоящим и предсказанным классам, а также F1-мера на каждый класс отдельно.

4 Этап 3. Анализ полученных результатов

После полученных результатов был проведен ручной анализ примеров, которые сумели запутать модель. И был замечен часто повторяющийся паттерн: модель относила негативные отзывы в нейтральные. В частности было обнаружено, что среди негативных отзывов часто люди указывают не только негативную часть, но и положительные аспекты, то есть отзывы представляют из себя скорее смешанный сентимент. К примеру:

[1] https://huggingface.co/docs/transformers/model_doc/roberta

> Качество ткани *плохое*. В плечах **сидит хорошо**, при этом в талии висит как болохон. Шло чуть больше месяца.

> Это наверное *самый плохой* Хилтон из сети! *Старые* номера, *не особо приветливый* персонал, одеяла и подушки не фирменные. Единственный плюс - это **очень красивая** территория вокруг отеля, **отлично подходит** для прогулок!!!

Это интересно, поскольку авторы оригинальной статьи наблюдали такие примеры в нейтральной категории, что кажется более логичным, чем наличие подобных отзывов в категории негативных. Таким образом, проанализировав получившиеся результаты, мы решили провести переразметку негативных отзывов, на которых ошиблась модель, аналогично оригинальной процедуре переразметки нейтральных отзывов в статье. Мы сделали дополнительную категорию для смешанных отзывов (mixed) и отнесли к ней примеры, в которых было выражено несколько тональностей. В отличие от авторов статьи, из-за недостатка ресурсов, мы решили не переразмечать все примеры вручную, а переразметить вручную только примеры из тестового датасета, на которых ошиблась модель, дообучить на этой переразметке RoBERTa и с помощью неё переразметить негативные отзывы из обучающего датасета, на которых ошиблась модель. К тому же, нам показалось интересной задача дообучения модели для переразметки подобных отзывов. Это в перспективе может быть полезно для работы с сентимент-бенчмарками на русском языке, в которых наблюдаются похожие закономерности

5 Этап 4. Обучение маленькой модели

На данном этапе планировалось дообучить модель на размеченных данных, чтобы переразметить весь датасет. К величайшему сожалению, мы не учли, что дисбаланс данных получается слишком сильный. Из-за этого наша модель дообучилась некорректно, так, например, на отрицательном сентименте у нас получился $f1 = 0$. После чего было принято решение обучить модель бинарной классификации на mixed и negative и переразметить с помощью неё только негативные примеры из оригинального датасета. На переразмеченных данных была обучена модель, показавшая следующие результаты см. Таблица. 2.

	negative	mixed
F1	0.792	0.896

Таблица 2: Иллюстрация обучения модели бинарной классификации.

На данный момент *в работе* находится этап переразметки всех негативных отзывов датасета, после чего планируется провести и следующий (финальный) этап цикла динамичного датасета: как на этапе 2 файн-тьюнить RoBERTa на уже новом датасете. Полученные результаты будут представлены на защите.

Список литературы

[1] Liu Y. et al. «Roberta: A robustly optimized bert pretraining approach». B: *//arXiv preprint arXiv:1907.11692.* (2019).

[2] Potts C. et al. «DynaSent: A dynamic benchmark for sentiment analysis». B: *///arXiv preprint arXiv:2012.15349.* (2020).

[3] Malafeev A. Rybakov V. «Aspect-Based Sentiment Analysis of Russian Hotel.» B: *ceur-ws.org/Vol-2268/paper8* ().

[4] Komarov M. Smetanin S. «Sentiment analysis of product reviews in Russian using convolutional neural networks.» B: *//2019 IEEE 21st conference on business informatics (CBI). – IEEE, 2019. – T. 1. – C. 482-486.* (2019).