Отчет по выполнению задания

по реализации модели для определения класса сообщения

06.02.2021.

Изучение текста задания.

Ознакомление с репозиторием intelligent-algorithm-for-automated-processing-of-citizenss-requests-master автор Андрей Петров.

Ознакомление с репозиторием classification-appeals-ada-master автор Кварацхелия Александр Гонерович.

Подготовка файла «NLP-preparation.ipynb» с предобработкой датасета заявок региона Татарстан. Заявки нужно классифицировать по 36 категориям. Предобработка взята из задания из летней практики по определению тональности твитов. Удаляем цифры и спецсимволы, замена «е» на «ё», замена смайлов на текст, удаление дублирования букв (более 2), удаление лишних пробелов. Лемматизация слов.

07.02.2021

Определение двух статистических моделей: логистическая регрессия и svm (метод опорных векторов). В качестве токенизатора использовал «мешок слов» и tf-idf.

Определение метрик качества модели: 'Accuracy', 'Precision', 'Recall', 'f1'.

Метрики вычислил с помощью кросс-валидации, разделив всю обучающую выборку на 5 частей и взяв среднее значение метрик для каждой части.

При обучении моделей были найдены оптимальные параметры регуляризации.

В таблице 1 приведены результаты обучения

Таблица 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Accuracy | Precision | Recall | f1 |
| Мешок слов + логистическая регрессия | 0.801 | 0.802 | 0.801 | 0.796 |
| Tf-idf + логистическая регрессия | 0.805 | 0.807 | 0.805 | 0.800 |
| Tf-idf + svm | 0.802 | 0.802 | 0.802 | 0.794 |

Полученные результаты на самом деле удивительны, так как количество классов велико, а с увеличением количества классов следовало бы ожидать ухудшение результатов классификации при критерии «один-ко-многим».

9.02.2021

Обучение нейросетевой модели. В качестве токенизатора использовал словарь из обученной word2vec модели (взял 600 000 слов из этого словаря). Модель word2vec была обучена на твитах в рамках летней практики в «интерсвязи». Также использовал статью

Использовал топологию сверточной нейронной сети. В первом слое использовал информацию о семантической близости слов, заложенную в word2vec модели (слой embedding). Нейросеть обучалась 200 эпох с замороженным embedding-слоем и 20 эпох – с размороженным.

На рисунке 1 показаны значения метрик качества модели в процессе обучения 200 эпох на обучающей и валидационной выборке.

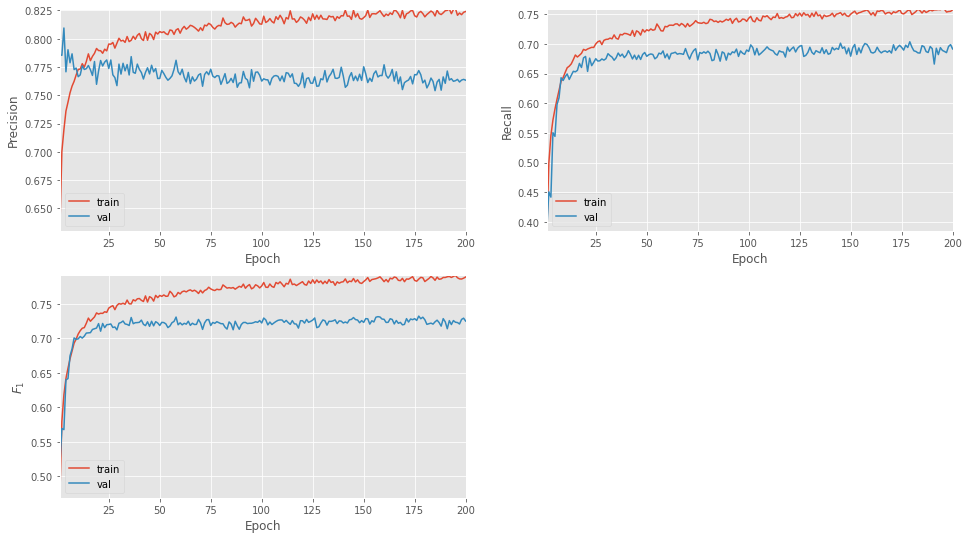
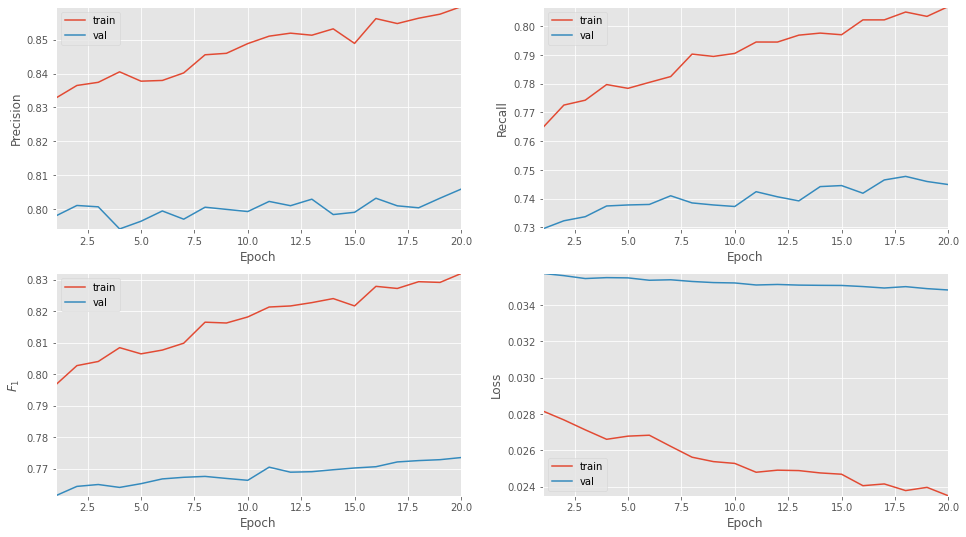


Рисунок 1

На рисунке 2 показаны значения метрик при дообучении модели



Получившиеся значения на тестовом наборе представлены в таблице 2.

Таблица 2

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Precision | Recall | f1 |
| w2v + CNN | 0.779 | 0.756 | 0.767 |

Здесь приведены микро-средние значения метрик.

Макро-среднее вычислит метрику независимо для каждого класса, а затем возьмет среднее (следовательно, будет обрабатывать все классы одинаково), тогда как микро-среднее будет агрегировать вклады всех классов для вычисления средней метрики. В мультиклассовой классификации предпочтение отдается микро-среднему, если вы подозреваете, что может быть дисбаланс классов.

16.02.2021

Обучение модели LSTM (файл NLP\_word2vec-LSTM\_2.ipynb).

Добавлен LSTM-слой в модель.

Лучшие результаты (микро-средние значения метрик) после 200 эпох обучения приведены в таблице 3.

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Precision | Recall | f1 |
| w2v + LSTM | 0.787 | 0.763 | 0.775 |

Таблица 3

27.02.2021

Применен токенизатор из предобученной BERT модели (файл NLP\_bert\_tokenizer\_2.ipynb)

Здесь возникли определенные сложности с использованием API модели. По времени, ушло 3 вечера. Разбирался по статье с Хабра <https://m.habr.com/ru/company/sberdevices/blog/527576/>, оттуда же взята руссокоязычная модель BERT для tf-hub.

В таблице 4 приведены лучшие результаты (микро-средние значения метрик) после 200 эпох обучения.

Таблица 4

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Precision | Recall | f1 |
| bert\_tokenizer + CNN | 0.920 | 0.914 | 0.917 |

Таким образом, на мой взгляд токенизация (входное представление) слов играет важную роль при подготовке нейросетевых и статистических моделей.