1. Описание проекта

На основе данных текстов объявлений об аренде квартир с Avito была разработана модель, выявляющая некоторые особенности объявлений, следующие из их текстового описания (например, арендодатель сдает квартиру только гражданам РФ или только девушкам). Этапы построения модели включают в себя:

1. Парсинг текстов объявлений и их разметку;
2. Предобработку данных;
3. Выбор архитектуры модели;
4. Оценку качества работы модели на тестовых данных.

1.1. Постановка задачи

Более конкретно, модель должна выделять следующие особенности объявлений:

1. Наличие предпочтений к национальности/внешности арендатора (сдается только славянам, только гражданам РФ и др.)
2. Наличие предпочтений семьям;
3. Наличие предпочтений к полу арендатора;
4. Наличие ограничений на максимальное количество проживающих (возможность классификации таких объявлений в демонстрационном приложении временно отключена).

Такая задача в общем случае может решаться как задача multilabel классификации, но в данной работе она решается как 4 отдельные задачи бинарной классификации (из-за малого числа возможных классов).

1.2. Получение и предобработка данных

Исходные тексты для обучения модели были получены с Avito путем парсинга с использованием библиотеки BeautifulSoup.

Метки к текстам проставлялись вручную. Исходные данные хранятся в csv-файле со столбцом текстом и четырьмя столбцами меток со значениями 0/1.

После получения данных производится их предобработка: токенизация, лемматизация с использованием PyMorphy2 и удаление стоп-слов.

Архитектура	Test ROC AUC/F1 score			
	Национальность	*Семьи*	*Пол*	*Ограничение на число проживающих*
Однослойная CNN	0,985/0,909	0,979/0,783	0,939/0,503	0,92/0,52
Двуслойная CNN	0,979/0,889	0,971/0,773	0,835/0,340	0,891/0,48
Однослойная RNN	0,929/0,832	0,98/0,804	0,659/0,07	0,923/0,679
Двуслойная RNN	0,982/0,936	0,979/0,784	0,695/0,17	0,850/0,556
Двуслойная RNN bidirectional	0,96/0,853	0,98/0,788	0,843/0,51	**0,960/0,671**
RuDBert	**0,991/0,942**	**0,984/0,79**	**0,988/0,764**	0,969/0,60

Рис. 1.1. Показатели ROC AUC и F-score для различных архитектур

1.3. Выбор модели и оценка ее качества

В работе был рассмотрен ряд нейронных сетей в качестве классификаторов:

1. Сверточные нейронные сети (однослойные и двуслойные)
2. Рекуррентные сети;
3. Использование модели (DistilRuBert) - уменьшенной версии Bert-модели, обученной на русскоязычном корпусе текстов.

Оценка качества работы моделей была произведена путем измерения среднего максимального F-score при проведении кросс-валидации на 5 фолдов. Показатели ROC AUC и F-score представлены на рис. 1.1.

Как видно из рисунка, наилучшие результаты при выявлении ограничений на максимальное число проживающих дает использование двунаправленной рекуррентной нейронной сети. При классификации по другим классам лучше всего работает DistilRuBert. Именно эти архитектуры и будут использоваться для классификации по соответствующим классам.

После обучения модели выбираются пороги классификации, обеспечивающие желаемое соотношение precision и recall (подробнее в notebooks/Thresholds Exploring.ipynb).