Эффективность классификаторов назначения платежа

Э.С. Зубчук[1,2], Д.И. Меньшиков[2], Н.Э. Михайловский[1,2]

[1]*Высшая IT-школа, Томский государственный университет, Ленина, 36, Томск, Россия, 634050*
[2]*NTR Labs, 12-ый проезд Марьиной Рощи 9с1, Москва, Россия, 127521*

Аннотация
Традиционно Центральный Банк России в рамках надзорной деятельности для классификации платежей использовал регулярные выражения. Они часто занимают несколько страниц, чтобы охватить необходимые ключевые слова и их словоформы. Мы сравнили этот метод с двумя современными подходами к классификации текстов – fastText и BERT с точки зрения скорости и точности.

Ключевые слова
Классификация коротких текстов, KYC, fastText, BERT

1. Обзор существующих решений

Поле “назначение платежа” - короткий текст максимальной длиной 210 символов. Различные исследования ([1], [2], [3]) ранее отмечали следующие особенности коротких текстов:

- содержат 5-30 слов, поэтому не обеспечивает достаточно контекста для традиционных подходов к класификации;
- включают множество орфографических ошибок, нестандартных лексем и шума;
- распределение классов несбалансированно;
- невозможность вручную разметить большой обучающий набор.

Большинство традиционных методов (таких как SVM [4], BAYES и kNN), основанных на сходствах частотных признаков, не учитывают вышеописанные особенности коротких текстов [3]. Несмотря на удовлетворительные результаты в решении различных задач классификации, они не подходят для классификации коротких текстов в силу их природы, сложности и проблем, связанных со спецификой, описанной выше. Таким образом, выбор представлений, вложенности, признаков, эффективное уменьшение размерности задачи и шума, а также повышение точности становятся проблемой в задаче классификации коротких текстов.

2. Данные

Мы работали с набором данных из ~600.000 назначений платежа. Все платежи были разбиты на 5 классов (продукты питания, строительные материалы, монтажные работы, табачная продукция, и “другое") с помощью регулярных выражений. Мы убрали все небуквенные символы и стоп-слова. Таким образом, платеж "Оплата по накл. №68659 от 24.12.2018 за табачные изделия Договор поставки №204-17 от 03.08.2017(частично) Сумма 34458-79 в т.ч. НДС 18% - 5256.47" преобразуется в 'накл табачные изделия поставки частично т ч'. Данные были сбалансированы и разделены на обучающую и тестовую выборку в соотношении 4:1. Так как датасет был размечен регулярными выражениями, уровень шума был для нас серьезной проблемой. Например, из-за большого количества общих терминов, смешение классов 'строительные материалы' и 'монтажные работы' было высоким. Обученные модели продемонстрировали высокий уровень согласованности с разметкой регулярными выражениями и частично исправили неточности, несмотря на несовершенство данных.

3. Модели и метрика качества

Мы сравнили с регулярными выражениями следующие модели: TF-IDF + наивный Байес, трансформер (ruBERT) и fastText. Точность моделей оценивалась средним взвешенным F1-score.

Важно отметить, что приведенные ниже метрики отражают качество модели относительно разметки с помощью регулярных выражений. Таблица 1 содержит показатели F1-score для каждой модели и класса.

Таблица 1
Показатели F1-score для каждой модели и класса

F1, %	TF-IDF + Bayes	ruBERT	fastText
Продукты питания	95	99	97
Строительные материалы	90	98	94
Монтажные работы	85	95	96
Табачная продукция	0	100	97
Другое	97	99	91
Среднее	**73**	**98**	**95**
Среднее взвешенное	**94**	**99**	**96**

3.1. Производительность

TF/IDF и FastText нетребовательны к системным ресурсам, а обучение занимает секунды. Классификатор RuBERT для обучения на нашем датасете требует графическую карту с 6+ Гб видеопамяти и 4 часа файнтюнинга на видеокарте Tesla T4.

3.2. Анализ результатов

Как видно из приведенных выше результатов, fastText и ruBERT классификаторы демонстрируют близкие показатели точности, при том, что ruBERT примерно в 1000 раз медленнее. Примечательно, что оба успешно обнаруживают табачные изделия, несмотря на крайне низкую представленность класса в датасете.

4. Заключение

Учитывая производительность и точность, модель на основе fastText была рекомендована Центральному Банку России в качестве замены регулярных выражений.

Точность модели при коммерческом использовании может быть несколько ниже из-за терминов, не встречающихся в обучающем датасете. Можно улучшить качество моделей, добавив синтетические данные, основанные на словарях, для каждой категории, кроме “другое”. Кроме того, обучающий датасет может быть нормализован так, чтобы экземпляры каждого класса были сбалансированы. Для повышения производительности модели может быть использована итеративная псевдоразметка для дополнительного обучения в условиях ограниченных размеченных данных и высокого уровня шума [5].

5. Литература

[1] Rui, Y. Dynamic Assembly Classification Algorithm for Short Text / Y. Rui, C. Xian-bin, L. Kai // Acta Electronica Sinica. – 2009. – Vol. 37(5). – P. 1019-1024.

[2] Jin-shu, S. Advances in Machine Learning Based Text Categorization / S. Jin-shu, Zh. Bo-feng, X. Xin // Journal of Software. – 2006. – Vol. 17(9). – P. 1848-1859.

[3] Song, G. Short Text Classification: A Survey / G. Song, Y. Ye, X. Du, X. Huang, S. Bie // J. Multimed. – 2014. – Vol. 9(5). – P. 635-643. DOI: 10.4304/jmm.9.5.

[4] Вапник, В.Н. Узнавание образов при помощи обобщенных портретов / В.Н. Вапник, А.Я. Лернер // Автомат. и телемех. – 1963. – Vol. 24(6). – P. 774-780.

[5] Arazo, E. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning / E. Arazo, D. Ortego, P. Albert, N.E. O’Connor, K. McGuinness [Electronic resource]. – URL: https://arxiv.org/abs/1908.02983.