Выявление метафор в текстах при помощи машинного обучения

К.ф.н., доцент кафедры ТиПЛ РГФ ВГУ Донина Ольга Валерьевна

Создание классификатора

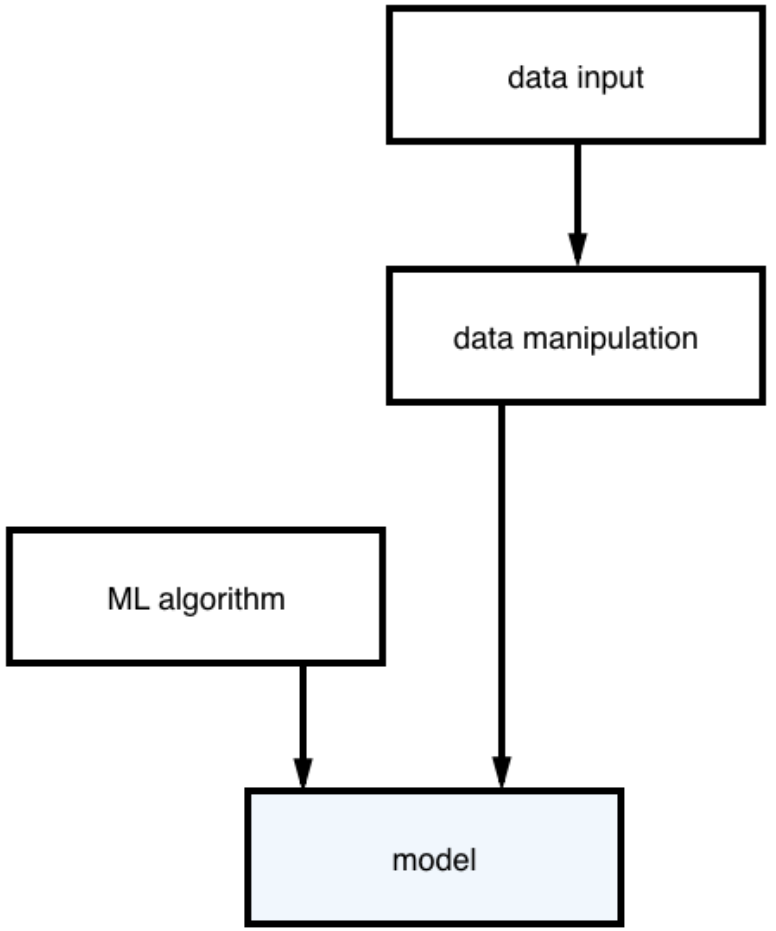
Варианты:

- Поиск по шаблонам,
- Использование грамматик,
- Машинное обучение (и возможное последующее использование XAI),
- Кластеризация имеющихся примеров и выявление отличительных характеристик для последующего использования их классификатором,
- Синтаксический парсинг и классификация с учетом его данных.

Создание классификатора

Варианты:

- Поиск по шаблонам,
- Использование грамматик,
- Машинное обучение (с учителем),
- Кластеризация имеющихся примеров и выявление отличительных характеристик для последующего использования их классификатором,
- Синтаксический парсинг и классификация с учетом его данных.



Датасет = 167 766 размеченных примеров

Бинарная классификация по предложению

Библиотеки:

- scikit-learn (для векторизации)
- NLTK (для лемматизации)

Наивный Байес:

- Accuracy (MultinomialNB) = 0,7926777133671916.

Логистическая регрессия

Эксперименты:

- Разбиение по словам
- Разбиение по символам
- Удаление стоп-слов
- Стемминг
- Лемматизация

Лучшие результаты:

Логистическая регрессия

(3-char-gram, со стоп-словами, без лемматизации/стемминга)

	Precision	Recall	f1-score
0	0,87	0,90	0,89
1	0,84	0,79	0,81
Accuracy			**0,86**
Macro avg	0,85	0,84	**0,85**
Weighted avg	0,86	0,86	0,86

Нейронные сети

Библиотеки:

- Keras (нейронные сети)
- scikit-learn (векторизация)

Последовательные сверточные нейронные сети (CNN)

Лучшие результаты:

Сверточные нейронные сети

(эпохи = 10, слои = 6 (в том числе 2 слоя dropout), train – 70% данных, test – 30% данных, векторизация = 2 и 3 символа)

	Precision	Recall	f1-score
0	0,88	0,92	0,90
1	0,87	0,81	0,84
Accuracy			**0,88**
Macro avg	0,88	0,87	**0,87**
Weighted avg	0,88	0,88	0,88

Планы

- улучшить классификатор,
- развернуть загрузчик и классификатор на сервере ВГУ
- разработать классификатор для автоматического определения варианта английского языка и типа текста
- адаптировать под другие языки

Q & A

к.ф.н., доцент кафедры ТиПЛ РГФ ВГУ Донина Ольга Валерьевна

olga-donina@mail.ru