Использование синтаксических признаков для анализа тональности

Кострыкина Екатерина Витальевна

Студентка

ФиКЛ НИУ ВШЭ, Москва, Россия

E-mail: evkostrykina@edu.hse.ru

Научный руководитель *— Дубнов Юрий Андреевич*

На сегодняшний день существует множество различных подходов к векторизации текстовых данных. В большинстве методов игнорируется порядок слов в предложении. В данной работе был рассмотрен подход с использованием синтаксических конструкций в качестве признаков для машинного обучения в задаче анализа тональности.

Целью исследования была проверка гипотезы о том, что при использовании синтаксических биграмм, построенных по полному дереву зависимостей, в качестве признаков точность классификации улучшается. Для экспериментов использовались реальные данные отзывов к кинофильмам на английском языке, а в качестве метода машинного обучения были апробированы несколько отличающихся классификаторов.

В работе [1], было предложено использование, в качестве признаков, синтаксических поддеревьев, то есть деревьев, полученных при удалении некоторых вершин и связей из полного дерева зависимостей. Для решения задачи был применен SVM классификатор. В ходе проведения множества экспериментов было выявлено, что при использовании униграмм+биграмм+синтаксических поддеревьев качество классификации повышается.

Синтаксические связи — формально-семантические взаимоотношения между компонентами словосочетания. Сохранение информации о данных отношениях способствует приросту в качестве при решении задачи анализа тональности. Для представления данных в виде матрицы термин-документ, оригинальные предложения были представлены в виде синтаксических деревьев зависимостей. Далее связанные слова были объединены в одно в порядке главное + зависимое и приведены к векторному виду путем вычисления весов с помощью TF–IDF меры.

В задаче классификации текстов мы имеем десятки, а иногда и сотни тысяч признаков. В связи с этим, для обучения алгоритма имеющихся вычислительных ресурсов может быть недостаточно. Протестировав различные методы сжатия данных было решено оста-

Таблица 1: Результаты

Multinomial Naive Bayes

	до SVD			после SVD		
	precision	recall	F-score	precision	recall	F-score
униграммы	85.44	74.6	79.52	81.25	80.5	80.7
синт.бигр.	84.84	83.0	83.83	81.27	81.1	81.09
уни.+синт.	**82.08**	**87.0**	**84.39**	82.57	80.7	81.53

Logistic Regression

	до SVD			после SVD		
	precision	recall	F-score	precision	recall	F-score
униграммы	83.36	82.5	82.88	83.35	82.7	82.97
синт.бигр.	82.13	82.2	82.09	83.62	84.4	83.9
уни.+синт.	82.29	77.5	79.79	**85.47**	**83.4**	**84.38**

Random Forest

	до SVD			после SVD		
	precision	recall	F-score	precision	recall	F-score
униграммы	84.13	81.1	82.52	79.89	80.8	80.31
синт.бигр.	78.39	76.5	77.38	77.76	74.8	76.16
уни.+синт.	**85.4**	**81.2**	**83.17**	80.85	80.7	80.73

новиться на методе главных компонент, а точнее его модификации через сингулярные разложения (SVD).

В результате данного исследования было выявлено, что использование синтаксических биграмм, построенных по полному дереву зависимостей, действительно может повысить качество классификации. При использовании трех различных классификаторов лучшие показатели точности были достигнуты при комбинации униграмм и синтаксических биграмм. Использование метода понижения размерности заметно улучшило качество классификации логистической регрессии. Все результаты, полученные при проведении экспериментов, являются строго воспроизводимыми.

Литература

1. Matsumoto S., Takamura H., Okumura M. Sentiment Classification Using Word Sub-sequences and Dependency Subtrees. // In: Ho T.B., Cheung D., Liu H. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD, 2005.