Анализ эмоциональной окраски сообщений в микроблогах с помощью вероятностных моделей

Лебедева Екатерина Андреевна,
545 группа

руководитель: Луцив Д.В.
рецензент: Тузова Е.А.

Оценка эмоциональной окраски текста

- Первое упоминание в статьях 2002 года
- Применение в маркетинге, политологии и т.д.
- Является задачей классификации
- Может быть решена методами машинного обучения: с учителем или без учителя
- Методы обучения с учителем: метод опорных векторов, логистическая регрессия, наивный байесовский классификатор

Постановка задачи

- Проанализировать особенности задачи для микроблогов
- Сравнить базовые методы обучения с учителем для данных из микроблогов и выбрать лучший
- Предложить, обосновать и реализовать новый метод на основе выбранного
- Оценить результаты работы нового метода

Особенности задачи

- Короткие тексты
- Социальные взаимодействия
- Смайлы :-) :-(;-)  
- Сокращения NP = no problem, AKA = also known as
- Пролонгирования noooooooo, so cuuuuuute
- Хештеги #nlp, #sentiment

Сравнение методов обучения с учителем

	Наивный байесовский классификатор			Метод опорных векторов			Логистическая регрессия		
Метка	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
-1	0.82	0.75	0.78	0.86	0.73	0.79	0.87	0.71	0.78
1	0.74	0.82	0.78	0.74	0.87	0.80	0.73	0.88	0.80
среднее	0.79	0.78	**0.78**	0.80	0.80	**0.79**	0.80	0.79	**0.79**
обучение	**1 сек**			**750 сек**			**437 сек**		

Обучающая выборка: 1 млн. примеров.
Тестовая выборка: 204 “-”, 182 “+”примеров.

Наивный байесовский классификатор

Вычисление класса *c для неизвестного примера *t***

априорная вероятность
попадания примера в класс *c*

$$c^* = \arg\max_{c \in \mathcal{C}} [P(c)P(t \mid c)]$$

$$P(t \mid c) = \sum_{f \in \mathcal{F}} P(f \mid c)^{f(t)}$$

вероятность признака *f* при
условии класса *c*

количественная
характеристика признака *f*
в сообщении *t*

Изменение метода

- Байесовский наивный байесовский классификатор
- Продолжение функции на множество всех слов
- n-граммы/онтологии
- Условные вероятности для социальных взаимодействий

Иллюстрация

A good task manager makes getting stuff done so satisfying. Read why I like Wunderlist & other apps I’m loving http://shar.es/SCKi9

a good task manager makes getting stuff done so satisfying. read why i like **task management software** & other apps i’m loving **url**

good **tas**k **man**ag **mak**e **get** **stu**ff **don**e **so** sat**isf**yin**g** rea**d** whi
like **tas**k **man**ag **sof**twa**r** oth**er** app **lov**e

c	-1	**1**
P(t,c)	0.0884	**0.9116**

0,006 0,00001 0,000002 0,001 0,004 0,0005 0,0008 0,0005 0,000007 0,0004 0,00005

0,004 0,00001 0,000002 0,00001 0,00005 0,00009 **0,005**

Оценка метода

Наивный байесовский классификатор

Метка	Precision	Recall	F1-score
-1	0.82	0.75	0.78
1	0.74	0.82	0.78
среднее	0.79	0.78	**0.78**
обучение	**1 сек**		

Байесовский наивный байесовский классификатор с онтологиями и триграммами

Метка	Precision	Recall	F1-score
-1	0.88	0.74	0.80
1	0.75	0.88	0.81
среднее	0.82	0.81	**0.81**
обучение	**3 сек**		

Обучающая выборка — 1млн. примеров, тестовая — 386 примеров.

Результаты

- Проанализированы особенности задачи анализа мнений для микроблогов
- Проведено сравнение базовых методов: метод опорных векторов, логистическая регрессия, наивный байесовский классификатор
- Предложен и реализован метод на основе наивного байесовского классификатора
- Собран корпус для оценки, метод дал улучшение на 3%