Улучшение качества тональной классификации с использованием лексиконов

Русначенко Н.Л.
kolyarus@yandex.ru

1. Задача

- Построение модели на основе методов ML для задачи **тональной классификации сообщений сети *Twitter*.** (соревнования **SentiRuEval**)
- **Тональность определяется** для сообщения в целом, и по отношению к рассматриваемым в сообщении организациям:
- Задача решается отдельно для каждой организации (коллекции данных):
  - **BANK** – банковские компании;
  - **TKK** – телекоммуникационные компании.
- Сообщению может быть проставлена одна из следующих тональных оценок: {1, 0, -1}

2. Идея

- Использование признаков на основе лексиконов словарей, состоящих из пар (t, v), где $t -$ терм, $v \in \mathbb{R}$ – тональная окраска терма.
- Увеличение объема обучающей коллекции (авторазметка и пополнение сообщениями внешних источников);

3. Смежные работы

- Автообучение: авторазметка сообщений с целью дополнения тональных классов обучающих коллекций **(Severyn A., Moshitti A., 2015)**
- Построение лексиконов на основе:
  - **PMI** – *меры взаимной информации*
  - **SO** – *сематической ориентации*

  (Turney P., 2002)
- Использование вспомогательных признаков, в т.ч. на основе лексиконов **(Saif. M. Kiritchenko S., Xiaodan Z., 2015)**

4. Построение лексиконов

На основе меры **взаимной информации** термов t_1, t_2**:**

$$\boldsymbol{PMI}(t_1, t_2) = \log_2 \frac{P(t_1 \wedge t_2)}{P(t_1) \cdot P(t_2)}$$

Введем **маркер** в качестве одного из параметров PMI. Возможные значения:

- `Excellent` -- хороший
- `Poor` -- плохой

Семантической ориентацией, называется величина:

$$\boldsymbol{SO}(t) = PMI(t, \texttt{Excellent}) - PMI(t, \texttt{Poor})$$

- Знак $SO(t)$ – определяет один из двух маркеров, к которому принадлежит t
- $|SO(t)|$ – степень принадлежности маркеру.

Лексикон составляется на основе коллекции K**:**

$$S: \{ \langle t, SO(t) \rangle \mid t \in K_{Excellent} \vee K_{Poor} \}$$

- $K_{Excellent}$ -- сообщения с меткой `Excellent`.
- K_{Poor} -- сообщения с меткой `Poor`.

5. Автогенерация коллекций

- Прием трансляции сообщений сети ***Twitter***.
- Разбиение полученной коллекции сообщений K на $K_{Excellent}$ и K_{Poor} с помощью:
  - Эмотиконов в сообщении.

6. Построенные лексиконы

1. На основе корпуса коротких текстов **Ю. Рубцовой**;
2. Сообщений сети *Twitter* за **январь 2016 года**
3. Тональный словарь созданный вручную экспертами

№	$K_{Excellent}$ термов	K_{Poor} термов	Всего термов
1	62 637 (56%)	50 177 (44%)	112 814
2	7 370 (3%)	228 721 (97%)	236 091
3	2 774 (26%)	7 148 (67%)	10 668

7. Подход

Классификация *методом опорных векторов*, **SVM** (библиотека LibSVM, Python)

Обработка сообщений:

- Лемматизация сообщений (**Mystem**,Yandex)
- Удаление символов ‘RT’, @пользователей, URL (из метаинформации остаются #хэштеги).Используемая весовая мера *TF-IDF*;
- Использование стоп слов;
- Замена лемм на тональные префиксы‘**+**’,’**–**’:

Сейчас хорошо работать не то что раньше

Сейчас +работать –то что раньше.

Признаки классификации:

- Учет эмотиконов (смайликов ☺, ☹);
- Число слов записанных в верхнем регистре;
- Число подряд идущих знаков {'?', '...', '!'}.
- Вычисление суммы $x = \sum SO(t),\ t \in S$, термов t , составляющих сообщение и входящих в лексикон S.

8. Обучающие коллекции

- **Несбалансированные:**
  - Предоставленные организаторами

2015				
Коллекция	☺	😐	☹	всего
BANK	356 (7%)	3 482 **(71%)**	1 077 (21%)	4 915
ТКК	956 (19%)	2 269 **(47%)**	1 634 (34%)	4 859
2016				
BANK	1 354 (15%)	4 870 **(55.4%)**	2 550 (29%)	8 783
ТКК	704 (7%)	6 756 **(74.22%)**	1 741 (19%)	9 102

- **Сбалансированные**:
  - *Балансировка* на основе коллекции **Ю. Рубцовой** построен лексикон и произведен отбор сообщений $m = \{t_i\}_{i=1}^{N}$ из той же коллекции по формуле:

$$\max_{i=1..N} |SO(t_i)| > P$$

P – пороговое значение, t_i– термы сообщения.

- α – сбалансированная коллекция 2015.
- β – балансировка коллекций 2015 и 2016 (их объединений) годов.

Сбалансированные		
Коллекция	α	β
TTK	6888	14610 (+112%)
BANK	10446	20268 (+94%)

9. Результаты

Параметры прогонов:

- **№1**– только русскоязычные термы и хэштеги;
- **№2** – №1 + применение тональных префиксов, использование лексиконов 1 и 2, учет всех признаков;
- **№3** – №2 + использование всех лексиконов.

Обучение на коллекции β **показывает прирост оценки** (правый столбец).

Мера оценки качества: $\boldsymbol{F_1 macro_{(neg,pos)}}$

BANK (SentiRuEval-2016)		
№	α	β
1	0.384	0.4536 **(+18.1%)**
2	0.3849	0.4672 **(+20.9%)**
3	**0.3862**	**0.4683** **(+21.25%)**

TKK (SentiRuEval-2016)		
№	2016	β
1	0.4849	0.5103 **(+5,2%)**
2	0.4832	0.5231 **(+8.2%)**
3	**0.5099**	**0.5286** **(+3.6%)**

10. Улучшение

- b – **baseline** результаты, относительно которых отмечается изменение качества.
- Настройка параметра C штрафной функции SVM классификатора (влияет на размер отступа разделяющей гиперплоскости):

$$C = 0.5$$

Улучшенные результаты, C = 0.5		
№	BANK	TKK
b	0.4536	0.5103
1	0.4558 **(+0.48)**	0.5235 **(+2,58%)**
2	**0.4795** **(+5.70)**	0.5338 **(+4,60%)**
3	0.4768 **(+5.11)**	**0.5452** **(+6,83%)**

- Добавление новых признаков y, z: вычисление min и max значений (с учетом нормализации) среди всех термов t_i сообщения m по каждому из лексиконов.

$$y = \min_{i=1..N} SO(t_i),\ t_i \in m, t_i \in S$$

$$z = \max_{i=1..N} SO(t_i),\ t_i \in m, t_i \in S$$

Улучшенные результаты, C = 0.5, использование новых признаков		
№	BANK	TKK
	0.4795	0,5452
1	0.4955 **(+3.34%)**	0.5259 **(-3.53%)**
2	0.5012 **(+4.53%)**	0.5283 **(-3.09%)**
3	**0.5239** **(+9.52%)**	**0.5453** **(+0.01%)**

Вывод

- Стабильное повышение качества классификации.
- Наибольший прирост достигается для задачи BANK

Прирост качества	**BANK**	**TKK**
Общий	+36,4%	+12,4%

Возможные дальнейшие улучшения:

- Использование иерархической классификации;
- В вычисление признаков на основе лексиконов добавить зависимость от TF-IDF весов.