USE OF LEXICONS TO IMPROVE QUALITY OF SENTIMENT CLASSIFICATION

Rusnachenko N.L.
Bauman Moscow State University (BMSTU), Moscow
kolyarus@yandex.ru

1. Problem

- Building Machine Learning based model for **Twitter messages sentiment classification task**. (*SentiRuEval* competition)
- **Sentiment class defines** for whole message, and shows relationship between message and company mentioned in it.
- For each domain this problem resolves separately:
  - **BANK** – bank companies;
  - **TCC** – telecommunication companies.
- Each message could be labeled with one of the following scores: {**1**, **0**, **-1**}

2. Solution

- Use lexicon based features:
  - **Lexicon** – dictionary, which consist of pairs (t, v), where t – term, $v \in \mathbb{R}$ – sentiment score.
- Increasing volume of training collections:
  - Balancing sentiment classes;
  - Adding and labeling messages from external sources;

3. Used articles

- Building lexicons (the main idea):
  - **PMI** – *Pointwise mutual information;*
  - **SO** – *Semantic orientation*
    - (Turney P., 2002)
- On the **Automatic Learning** of Sentiment Lexicons, Human Language Technologies (Severyn A., Moshitti A., 2015)
- NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets (Saif. M. Kiritchenko S., Xiaodan Z., 2015)

4. Building Lexicons

Based on **pointwise mutual information** of terms t_1, t_2:

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

Introducing **marker** as a second parameter of PMI function. Possible marker values:

- `Excellent`;
- `Poor`.

Semantic orientation is a function:

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

- $sgn(SO(t))$ – determines the marker type of term t;
- $|SO(t)|$ – degree of belonging.

Building lexicon from messages of collection $K = K_{Excellent} \cup K_{Poor}$:

$$S: \{\ (t, SO(t))\ |\ t \in K_{Excellent} \cup K_{Poor}\}$$

- $K_{Excellent}$ -- messages labeled `Excellent`.
- K_{Poor} -- messages labeled `Poor`.

Making sentiment collections K from scratch (automatic labeling):

- Receive messages via Streaming *Twitter API, and* composing collection K.
- Split collection messages K with $\boldsymbol{K_{Excellent}}$ and $\boldsymbol{K_{Poor}}$ by means of:
  Message emoticons (`:-)`, `:-(`, `xD`, …);

5. Lexicons

1. Messages of **Yu. Rubtsova short message corpus**;
2. Twitter messages through the **January, 2016;**
3. Sentiment vocabulary **RuSentiLex**.

#	$K_{Excellent}$ Terms count	K_{Poor} Terms count	Total Terms count
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

6. Approach

Support Vector Machine **(SVM)** as a classifier**,** linear classification kernel.

Message processing:

1. Lemmatization (**Mystem**, Yandex);
2. Removing 'RT' symbols, @users, URL (message metainformation contains only #hashtags). Weight measure: *TF-IDF*;
3. Applying list of stop words;
4. Replacing predefined lemmas with sentiment prefixes '**+**','**–**':

Сейчас хорошо работать не то что раньше
Сейчас +работать –то что раньше.

Classification features:

- ✓ Emoticons ($\sum$ of positive and negative);
- ✓ Amount of UPPERCASE words;
- ✓ Amount of signs {'?', '...', '!'}.
- ✓ Calculating sum $x = \sum SO(t), t \in L$, of terms t composes message and exist in lexicon L.

7. Training Collections

- **Imbalanced collections:**
  - Provided by *SentiRuEval* organizers:

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

- **Balanced collections**:
  - *Balancing:* filtering messages $m = \{t_i\}_{i=1}^{N}$ from *Yu. Rubtsova corpus* (by means of Lexicon, based on the same corpus) by formula:

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

Bound – bounding value, t_i– message terms.

- $\boldsymbol{\alpha}$ – balanced 2015 train collection.
- $\boldsymbol{\beta}$ – united collections of 2015 and 2016 years, and then balanced.

Balanced collections (messages count)	α	β
BANK	10446	20268 (+94%)
TCC	6888	14610 (+112%)

8. Results

Features settings:

- **№1**– only Russian terms and hashtags;
- **№2** – №1 + using sentiment prefixes ('+', '-'), all features (using lexicons only #1 and #2);
- **№3** – №2 + using lexicon #3.

Quality measure: $\boldsymbol{F_1 macro}_{(\boldsymbol{neg,pos})}$

#	BANK (SentiRuEval-2016) α	β
1	38.40	45.36 **(+6.96)**
2	38.49	46.72 **(+8.23)**
3	**38.62**	**46.83 (+8.21)**

#	TCC (SentiRuEval-2016) 2016	β
1	48.49	51.03 **(+2.54)**
2	48.32	52.31 **(+3.99)**
3	**50.99**	**52.86 (+1.87)**

- Using β as a train collection **improves classification quality** (right column).

9. Improvements

- $\boldsymbol{b}$ – baseline.
- C – SVM penalty function parameter value (affects the margin for hyperplane between classes). Default $C = 1$:

$$C = 0.5$$

Use C = 0.5 #	BANK	TCC
b	45.36	51.03
1	45.58 **(+0.22)**	52.35 **(+1.32)**
2	**47.95 (+2.59)**	53.38 **(+2.35)**
3	47.68 **(+2.32)**	**54.52 (+3.49)**

- Adding new lexicon based features y, z: for lexicon L, for each terms t_i in message m calculating min and max values $SO(t_i)$ (normalized values):

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

C = 0.5, Adding new lexicon based features #	BANK	TCC
b	47.95	54.52
1	49.55 **(+1.60)**	52.59 **(-1.93)**
2	50.12 **(+2.17)**	52.83 **(-1.69)**
3	**52.39 (+4.44)**	**54.53 (+0.01)**

Conclusion

- Classification quality stable improves after using balanced collection and lexicon based features.

Quality improvement	BANK	TCC
Total	**+13.99**	**+6.03**

Future possible improvements:

- Using hierarchy classification model;
- Calculating lexicon based features, depending on *TF-IDF* terms weights.