Sentiment Analysis: My Experiments

Wang, Sida, and Christopher D. Manning. "Baselines and bigrams: Simple, good sentiment and topic classification." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume* 2. Association for Computational Linguistics, 2012.

김미숙

2015.11.25

Contents

- Overview
- Experiment Settings
- Experiment Results
- To dos

1. Overview

- Multinomial Naïve Bayes(MNB)
  - Likelihood of observing a x is given by

$$p(x|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{k_i}^{x_i}$$

- Support Vector Machine (SVM)

$$\min_{x^{(k)},w,b} w^T w + C \sum_i \max\left(0, 1 - y^{(i)}\left(w^T \hat{f}^{(i)} + b\right)\right)^2$$

- SVM with NB features (NBSVM)
  - Use $\tilde{f}^{(k)} = \hat{r} \circ \hat{f}^{(k)}$, instead of $x^{(k)} = \tilde{f}^{(k)}$

2. Experiment Settings

- **Data description**
  - The number of companies: 4(Financials)
    - C: Citigroup Inc
    - WFC: Wells Fargo & Co
    - GS: Goldman Sachs Group Inc
    - JPM: JP Morgan Chase & Co
  - The number of documents: 2,032
  - The number of matched documents: 1,810

2. Experiment Settings

- **Response Variable**
  - Price change: close price on the next day - close price before the report occurrence
    - UP: more than 3%
    - DOWN: less than 3%
    - STAY: between UP and DOWN
  - Total indexes
    - djia: Dow Jones Industrial Average
    - **gspc: S&P 500**
    - ixic: NASDAQ

2. Experiment Settings

- **Experimental setup**
  - Split at spaces
  - Not use stopwords, lexicons or other resources
  - Used parameter
    - $\alpha = 1, C = 1, \beta = 0.25$ for NBSVM
    - $C = 0.1$ for SVM
  - 10-fold cross-validation

3. Experiment Results

- **Price change after public announcement**

	1 days	7 days	3 days
MNB-Uni	62.33	60.75	59.89
MNB-Bi	64.91	64.34	62.66
SVM-Uni	63.41	62.32	
SVM-Bi	63.80	64.92	63.91
NBSVM-Uni	62.87	61.38	
NBSVM-Bi	65.67	65.41	63.73

공시가 뜨고 난 다음날 종가의 가격 변화의 예측 성능이 일주일 뒤의 가격 변화에 대한 예측 성능보다 더 좋음

3. Experiment Results

- **MNB is better at snippets**
  - These model(MNB, SVM, NBSVM) perform better than many sophisticated, rule-based methods.
  - With the only exception being MPQA, MNB performed better than SVM in all cases. → In contrast to their result that an SVM usually beats NM when it has more than 30-50 training cases, MNB is still better on snippets even with relatively large training sets(9k cases).

Method	RT-s	MPQA	CR	Subj.
MNB-uni	77.9	85.3	79.8	**92.6**
MNB-bi	**79.0**	**86.3**	80.0	**93.6**
SVM-uni	76.2	86.1	79.0	90.8
SVM-bi	77.7	**86.7**	80.8	91.7
NBSVM-uni	**78.1**	85.3	80.5	92.4
NBSVM-bi	**79.4**	**86.3**	**81.8**	**93.2**
RAE	76.8	85.7	–	–
RAE-pretrain	77.7	**86.4**	–	–
Voting-w/Rev.	63.1	81.7	74.2	–
Rule	62.9	81.8	74.3	–
BoF-noDic.	75.7	81.8	79.3	–
BoF-w/Rev.	76.4	84.1	**81.4**	–
Tree-CRF	77.3	86.1	**81.4**	–
BoWSVM	–	–	–	90.0

Tree-CRF: (Nakagawa et al., 2010)
RAE: Recursive Autoencoders (Socher et al., 2011).
RAE-pretrain: train on Wikipedia (Collobert and Weston, 2008).
"Voting" and "Rule": use a sentiment lexicon and hard-coded reversal rules.
"w/Rev": the polarities of phrases which have odd numbers of reversal phrases in their ancestors

4. To Dos

● 실험 계획

- Data 개수를 늘려서 실험하여 성능이 증가하는지 확인
  - 현재 document 개수 약 2,000개 → 10,000개로 증가
- 공시가 떴을 때, 공시 전날 close price와 1day ~ 7 days 후의 가격 change를 살펴 보고자 함
  - 추가적으로 시간에 따른 가격 추이를 확인 할 수 있으면 좋을 것 같음
- 하나의 기업, 혹은 하나의 sector에 대해서 document의 sentiment들이 어떤 방향으로 변하는지 그래프로 나타내어 위치 변화를 시각화 하는 실험
- 한계점: 각 공시가 대체로 길이가 길다. 따라서 sentence 단위로 구분하여 subjectivity를 갖는 문장들만을 사용하여 document의 sentiment를 찾는다. → Sentiment analysis와 subjectivity를 동시에 할 수 있는 연구로 이어질 수 있음

4. To Dos

- **추후 일정**
  - 11월 말: 다양한 실험 수행하여 연구의 contribution을 만들어 나가고자 함
  - 12월 중순: 논문 작성