Emotion Classification of Chinese Microblog Text via Fusion of BoW and eVector Feature Representations

Chengxin Li, Huimin Wu, Qin Jin

Multimedia Computing Lab

School of Information

Renmin University of China

Outline

- Background
- Baseline system
- Contrast system
- Analysis and Conclusion

Background

- Social media
  - Twitter, Weibo, …
- Applications
  - Businesses: marketing
  - Government: public opinion monitoring
- Technologies
  - Opinion mining
  - Sentiment analysis
- The evaluation of NLPCC 2014
  - Emotion analysis for Chinese Weibo texts

Brief Introduction of the Task

- Emotion detection
  - Binary classification: whether has emotion
  - Single label: one label
- Emotion classification
  - Multi-class classification: if has emotion, classify it into corresponding emotion categories
  - Emotion categories: anger, disgust, fear, happiness, sadness, surprise
  - Primary emotion and secondary emotion

Baseline System

- Generate a dictionary
  - Top key words based on ***TF-IDF*** method
  - Special indicator in microblog texts
    - Punctuation repetition: “。。。”,“！！！”, …
    - Emoticons: [爱你]  [泪] 

      [抓狂]  [哈哈] 
- Represent a text using ***BoW*** method
- SVM classifier

Contrast System

- A new feature representation: ***eVector***
- Motivation
  - Represent a text with seven emotion dimensions directly instead of thousands of dimensions in ***BoW*** method
    - Generate a new dictionary for every emotion category
    - Represent a text using a vector composed of only seven dimensions

Words Types in Dictionary

- Intuition and observation: three types words in all the texts
  - Emotional word
    - “怒”(anger)
  - Common word
    - “真得”(really)
  - Not emotional & Uncommon word
    - “资本家”(capitalist)

Expected Distribution

Type	*n_i*	*n_o*	*n_l*
Emotional word	more	less	less
Common word	fair	more	more
Not emotional & Uncommon word	less	less	less

- *n_i*: occurrence times of one ***word*** in *emotion_i*
- *n_o*: occurrence times of this ***word*** in all the other emotion categories except *emotion_i*
- *n_l*: number of emotion categories this ***word*** appears in

Compute Weights for Words in Dictionary

$$weight = \frac{n\_i}{(n\_o * n\_l + 1)}$$

- Sort the words by weight in a descending order
  - Top
    - Emotional word
  - Middle
    - Not emotional & Uncommon word
  - Bottom
    - Common word

Part of the Examples for “anger” Emotion Category

Anger	*Word (translation)*	*weight*
Examples in the top part of the ranked list	恨死 (hate)	12.3
	气死我了(piss me off)	8.0
	MB (fuck)	7.0
	贱人 (bitch)	5.0
	这蛋 (bullshit)	5.0
Examples in the middle part of the ranked list	心肝儿 (darling)	0.5
	掩护 (cover)	0.5
	私车 (personal car)	0.5
	扭转 (reverse)	0.5
	秒钟 (clock)	0.5
Examples in the bottom part of the ranked list	挺 (very)	0.0031
	每 (every)	0.0029
	现场 (on site)	0.0029
	滴 (a drop)	0.0029
	害羞 (shy)	0.0029

New Representation of the Text

- ***eVector*** = $(d_1, d_2, d_3, d_4, d_5, d_6, d_7)$
  - Seven dimension corresponding to seven emotion categories
- $$d_i = \sum_{word_k \in emotion_i} weight(word_k)$$
  - *$word_k$ represents every word in this text*

多媒体计算实验室

Classification

- The first step
  - use the baseline system to perform emotion detection experiment
- The second step
  - use the baseline system and contrast system to do emotion classification experiment

Detection Results and Analysis

- Emotion detection results

	Precision	Recall	F-measure
Occurrence (D)	0.58	0.73	0.65
Frequency (D)	0.58	0.74	0.65
Occurrence (S)	0.58	0.50	0.54
Frequency (S)	0.57	0.52	0.56

- D: document level S: sentence level

- Analysis
  - Microblog text is too short
  - Top words occur in text only once

Classification Results and Analysis

- Emotion classification results

System	*Document Level*		*Sentence Level*	
	looseAP	*strictAP*	*looseAP*	*strictAP*
Baseline system (BoW)	0.44	0.40	0.33	0.31
Contrast system (eVector)	0.38	0.34	0.28	0.27
Fusion	**0.46**	**0.41**	**0.34**	**0.32**

- ***Fusion***: linear combination of baseline system and contrast system

Analysis

- ***eVector***: high weights for good indicators, large value for corresponding emotion dimension
  - Text has good indicators
    - 每天上班都是相同的路，真的想走不一样的路，有一天下很大的雨，要绕路走，结果那天迷了路，…,广州的路！[怒骂]
- ***BoW***: has more dimensions, comprehensive context information
  - Text has weak/no indicators
    - 如果南中国海的石油开采能让国内油价降低一分钱，我就支持维护南海主权，如果能降低一毛，我愿意多缴税做军费，如果降低一块，我愿意参军。如果只是维护三大石油集团的利益，跟我有毛关系呢？那些像打了鸡血一样的愤青，你坐飞机他们会给你燃油附加费打八折吗？
- ***Fusion***: complementary of ***BoW*** and ***eVector*** representations

Summarize

- Background of emotion analysis task
- A traditional representation: ***BoW***
- A novel representation: ***eVector***
- Some results and analysis