Sentiment Strength Detection for the Social Web论文笔记

Abstract

大多数情绪分析设计商业，如分析**产品评论**情感。但是**they exploit indirect indicators of sentiment that can reflect genre or topic instead**。这对于分析社交文本是不可取的。

无监督学习是SSth的优点。

数据集：**MySpace, Twitter, YouTube,Digg,RunnersWorld,BBC Forums**

结论：

本文评估了该算法的改进版本。SentiStrength优于原来版本，不优于机器学习方法。对于social web context表现足够好，对于情感弱的文本（新闻）表现不好。

Introduction

social web context只是情感分析的一小部分，但是涉及内容广泛，对社会学研究有意义。

不能使用传统情感分析方法分析social web context，因为：

1. 标注数据少。情感文本必须手动标注
2. 情感分析是领域依赖的。不同topics and communication styles需要不同的classifier。
3. 关键点：技术优化后的classifier可能会使用indirect indicators，会给出误导性结果。

   e.g.：政治话题中，`伊拉克、伊朗、巴勒斯坦和以色列`可能会成为消极词。因为它们通常出现在消极话题中。

所以我们使用direct indicators of sentiment进行情感分析。

方法：主要通过文本中存在的**已知的含情感词汇或短语的词汇**进行情感分析。

方法1：

  情绪分类：

  积极情绪强度；消极情绪强度。转化为分数。

方法2：

积极消极情绪合成单一指标。

研究目的：

SentiStrength作为MySpace sentiment strength detection program的升级版，只依赖direct indicator of sentiment的Lexical Algorithms。对于social web能否有效？

Sentiment Analysis

情绪分析任务：

- 主观性（subjective）检测

  subjective or objective

- 极性（polarity）检测

  positive or negative

- 情绪强度（sentiment strength）检测（不常用）

  predicts the **strength** of positive or negative sentiment

本节分析极性检测，三种检测方法类似。

常用情绪分析方法：

机器学习

- 选择机器学习算法
- 提取文本特征，训练classifier

种子词法

- co-occur frequency

缺点：使用indirect indicator

Lexical Algorithms

已知情绪集+算法 → 预测文本情绪

优点：

1. 可扩展其他信息。

   表情符号表、语义规则...

2. 词汇源丰富

3. 有各种方法改进标准源

   添加额外术语，已针对特定领域。如，“小”在便携设备中是积极词

SO-CAL

SO-CAL和本文的SentiStrength很相似。

negtive to positive

-5 to +5

词法建立

手工标注2,252 adjectives, 745 adverbs 1,142 nouns, and 903 verbs

强化情绪表达词，如would；重复字符；大写字符...

SentiStrength 2

定义：

SentiStrength is a **lexicon-based classifier** that uses additional (nonlexical) linguistic information and rules to **detect sentiment strength** in **short informal English text**.

输出：

two integers: 1 to 5 for positive sentiment strength; a separate score of 1 to 5 for negative sentiment strength.

e.g.:

1,1 不积极，不消极

3,5 适度积极，强烈消极

主要特征列表

本文P166

- A *sentiment word list with human polarity and strength judgements*^. Some words include Kleene star stemming (e.g., ador*).
  - The word "*miss*" is a special case with a positive and negative strength of 2. It is frequently used to express sadness and loves simultaneously.
- A *spelling correction algorithm* deletes repeated letters in a word when the letters are more frequently repeated than normal for English or, if a word is not found in an English dictionary, when deleting repeated letters creates a dictionary word (e.g., hellp -> help).
- A *booster word list* is used to strengthen or weaken the emotion of following sentiment words.
- An *idiom list*^ is used to identify the sentiment of a few common phrases. This overrides individual sentiment word strengths.
- A *negating word list*^ is used to invert following emotion words (skipping any intervening booster words).
- *At least two repeated letters* added to words give a strength boost sentiment words by 1. For instance, haaaappy is more positive than happy. Neutral words are given a positive sentiment strength of 2 instead.
- An *emoticon list with polarities* is used to identify additional sentiment.
- Sentences with *exclamation marks* have a minimum positive strength of 2, unless negative.
- *Repeated punctuation* with one or more exclamation marks boost the strength of the immediately preceding sentiment word by 1.
- *Negative sentiment is ignored in questions*^.

1→2

SentiStrength 2相对 1 扩展了什么?

- The sentiment word list was extended with negative General Inquirer terms with human-coded sentiment weights and Kleene star stemming. This increased the number of terms in the sentiment word list from 693 to 2,310.
- The sentiment word terms were tested against a dictionary to check for incorrectly matching words and derivative words that did not match. This resulted in many terms being converted to wildcards (e.g., to match –ness word variants) and some exclusions being added (e.g., amazon* added as an exclusion for amaz*, admiral* added as an exclusion for admir*). Exclusions were typically rare words matching common sentiment words but longer. SentiStrength was recoded to match the longest term if multiple terms matched. This increased the sentiment word list to 2,489 terms, 228 of which were neutral (strength 1), either as exclusions or as potential sentiment words that could be incorporated by the training stage. Most (1,364) terms had a Kleene star ending after this stage.
- Negating negative terms makes them neutral rather than positive (e.g., "I do not hate him" is not positive).
- The idiom list was extended with phrases indicating word senses for common sentiment words. For instance, "is like" has strength 1 (the minimum score on the positive scale, indicating neutral text) because "like" is a comparator after "is" rather than a positive term (strength 2). This is a simple alternative to part of speech tagging for the most important sentiment word contexts relevant to the algorithm scores.
- The special rule for negative sentiment in questions was removed.

Research Questions

研究目标

- 评估SentiStrength2在各种不同的在线环境中，是否可行。
- 评估SentiStrength2与其他使用indirect terms的方法、机器学习相比，表现如何。

Methods and Data

test data set 6+1

- BBC Forum posts: Public news-related discussions. This represents discussions about various serious topics, from national and world news to religion and politics.
- Digg.com posts: Public comments on news stories. This represents general news commentary and evaluation.
- MySpace comments: Public messages between Friends in this SNS. These data represent SNS communication.
- Runners World forum posts: Public group messages on the topic of marathon running. These data represent specialist forums for common-interest groups.
- Twitter posts: Public microblog broadcasts. Twitter is an important site in its own right.
- YouTube comments: Text comments posted to videos on the YouTube website. This represents comments on resources and any associated discussions.
- All six combined: All of the above were combined into a single large data set to assess how well SentiStrength 2 performed in a mixed environment and to see whether a significant increase in training data would give a large relative increase in the performance of the selected machine-learning methods.

- 证明测试数据有效

    - Krippendorff’s α
- 标准机器学习算法(已改进)
    - support vectormachines (SVM; Sequential Minimal Optimization variant, SMO)
    - Logistic Regression (SLOG for short)
    - ADA Boost
    - SVM Regression
    - Decision Table
    - Naïve Bayes
    - J48 classification tree, and JRip rule-based classififier

Corpus Statistics

数据集大小

TABLE 1. Text size statistics for each data set.

	Mean characters	Mean words	Texts
BBC	356.44	62.54	1000
Digg	183.32	31.49	1077
MySpace	101.91	20.08	1041
Runners World	335.42	65.13	1046
Twitter	94.55	15.35	4218
YouTube	91.18	17.12	3407
All six combined	146.05	26.18	11790

Overall Sentiment Distribution

见Figures 1 and 2

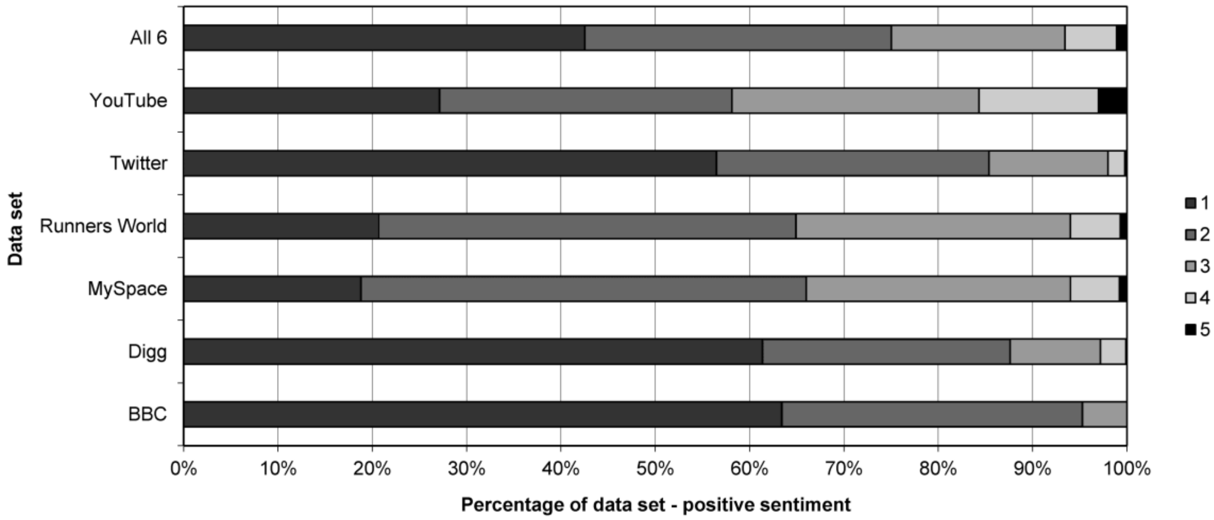

FIG. 1. The proportion of positive sentiment strengths in each data set.

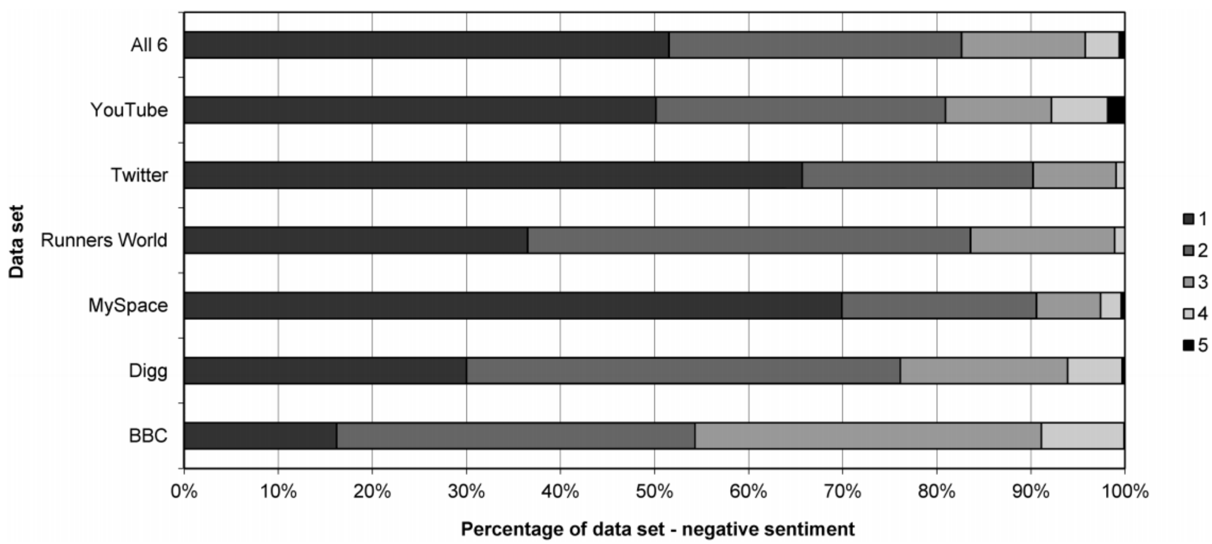

FIG. 2. The proportion of negative sentiment strengths in each data set.

Results

- correct：accuracy，测试集命中率
- +/- 1：accuracy，允许误差为 ±1class 之后的命中率
- correl：correlation，Pearson算法计算相关度

见 Table 2

TABLE 2. Unsupervised and supervised SentiStrength 2 against the baseline measure (predicting the most common class) and the standard machine learning algorithm and feature set size (from 100, 200 to 1000) having the highest correlation with the human-coded values.[a]

	+ve correct	−ve correct	+ve +/−1	−ve +/−1	+ve correl.	−ve correl.
BBC forums						
Baseline	63.4%	38.1%	95.3%	91.1%	–	–
Unsupervised ssth	51.3%	46.0%	90.3%	91.1%	0.296	**0.591**
Supervised ssth	60.9%	48.4%	94.5%	92.8%	0.286	0.573
	−0.2/+0.2	−0.3/+0.2	−0.1/+0.1	−0.1/+0.1	−4/+5	−3/+2
SLOG 200	**76.7%**		**97.2%**		**0.508**	
	−0.1/+0.1		−0/+0.1		−4/+4	
SLOG 100		**51.1%**		**94.7%**		0.519
		−0.2/+0.2		−0.1/+0.1		−3/+3
Digg						
Baseline	61.5%	46.1%	87.7%	**94.0%**	–	–
Unsupervised ssth	53.9%	46.7%	88.6%	90.8%	0.352	0.552
Supervised ssth	57.9%	50.5%	**92.0%**	92.9%	**0.380**	**0.569**
	−0.2/+0.2	−0.1/+0.2	−0.1/+0.1	−0.1/+0.1	−3/+3	−2/+1
SLOG 100	**63.1%**		90.9%		0.339	
	−0.2/+0.2		−0.1/+0		−7/+7	
SLOG 100		**55.2%**		93.6%		0.498
		−0.4/+0.3		−0.1/+0.2		−6/+6
MySpace						
Baseline	47.3%	69.9%	94.0%	90.6%	–	–
Unsupervised ssth	62.1%	70.9%	**97.8%**	**95.6%**	**0.647**	0.599
Supervised ssth	62.1%	72.4%	96.6%	95.3%	0.625	**0.615**
	−0.3/+0.2	−0.1/+0.2	−0/+0.1	−0.1/+0.1	−3/+3	−2/+3
SLOG 100	**63.0%**		96.8%		0.638	
	−0.2/+0.2		−0.1/+0.1		−2/+3	
SMO 100		**77.3%**		93.6%		0.563
		−0.1/+0.1		−0.1/+0.1		−5/+4
Runners World						
Baseline	44.2%	47.1%	94.0%	**98.9%**	–	–
Unsupervised ssth	53.5%	50.9%	94.7%	90.0%	0.567	0.541
Supervised ssth	53.9%	55.8%	**95.4%**	93.6%	0.593	0.537
	−0.3/+0.3	−0.3/+0.3	−0.1/+0.1	−0.1/+0.1	−2/+2	−2/+2
SLOG 200	**61.5%**		95.3%		**0.597**	
	−0.3/+0.3		−0.1/+0.1		−4/+4	
SLOG 300		**65.3%**		96.1%		**0.542**
		−0.2/+0.3		−0.1/+0.1		−4/+4
Twitter						
Baseline	56.5%	65.7%	85.4%	90.2%	–	–
Unsupervised ssth	59.2%	66.1%	94.2%	93.4%	0.541	0.499
Supervised ssth	63.7%	67.8%	94.8%	94.6%	0.548	0.480
	−0.1/+0	−0.1/+0.1	−0/+0	−0.1/+0	−2/+1	−2/+2
SLOG 200	**70.7%**		**94.9%**		**0.615**	
	−0.1/+0		−0.1/+0		−1/+1	
SLOG 200		**75.4%**		**94.9%**		**0.519**
		−0.1/+0.1		−0/+0.1		−2/+2
YouTube						
Baseline	31.0%	50.1%	84.3%	80.9%	–	–
Unsupervised ssth	44.3%	56.1%	88.2%	88.5%	0.589	0.521
Supervised ssth	46.5%	57.8%	**89.0%**	89.0%	0.621	0.541
	−0.2/+0.1	−0.1/+0.1	−0.1/+0	−0.1/+0	−1/+1	−1/+2
SLOG 200	**52.8%**		89.6%		**0.644**	
	−0.1/+0.1		−0/+0.1		−2/+1	
SLOG 300		**64.3%**		**90.8%**		**0.573**
		−0.1/+0.1		−0.1/+0		−3/+3
All 6						
Baseline	42.6%	51.5%	75.1%	82.7%	–	–
Unsupervised ssth	53.5%	58.8%	92.1%	91.5%	0.556	0.565
Supervised ssth	56.3%	61.7%	**92.6%**	**93.5%**	0.594	**0.573**
	−0/+0.1	−0.1/+0.1	−0.1/+0.1	−0/+0	−0/+1	−1/+0
SMO 800	**60.7%**		92.3%		**0.642**	
	−0/+0.1		−0/+0		−1/+1	
SMO 1000		**64.3%**		92.8%		0.547
		−0/+0.1		−0/+0		−1/+2

Note. Correlation is the most important metric.

[a]The metrics used are as follows: accuracy (% correct), accuracy within 1 (i.e. +/− 1 class), and correlation. Best values on each data set and each metric are in bold. When multiple tests are available then 30 are conducted and a 95% confidence interval is indicated underneath the mean. For instance, 60.9% above −0.2/+0.2 denotes a 95% confidence interval for the mean of (60.7%, 61.1%). For correlations, the confidence interval adjustments are for the third decimal place.

Limitations and Discussion

1. 数据集有限

2. 数据标注可能有误差

之后水了一堆ml的内容。

Conclusions

- 监督/非监督两个版本都表现良好，在全新的social web context中，无监督SSth可以有效检测情绪。
- SSth可能不如ml

SSth的无监督版本有应用价值。