Introduction to Sentiment strength detection

Mike Thelwall
University of Wolverhampton, UK
Sentiment Strength Detection with *SentiStrength*

1. Detect positive and negative sentiment *strength* in short informal text
   1. Develop workarounds for lack of standard grammar and spelling
   2. Harness emotion expression forms unique to MySpace or CMC (e.g., :-) or haaappppyyyy!!!)
   3. Classify simultaneously as positive 1-5 AND negative 1-5 sentiment

2. Apply to MySpace comments and social issues

SentiStrength 1 Algorithm - Core

List of 890 positive and negative sentiment terms and strengths (1 to 5), e.g.

- ache = -2, dislike = -3, hate=-4, excruciating -5
- encourage = 2, coolest = 3, lover = 4

Sentiment strength is highest in sentence; or highest sentence if multiple sentences
Examples

-2

-2

-2

You are the coolest.

3

3

3

I hate Paul but encourage him.

4

4

4

positive, negative

1, -2

3, -1

2, -4
Term Strength Optimisation

*Term strengths (e.g., ache = -2) initially fixed by a human coder*

*Term strengths optimised on training set with 10-fold cross-validation*

- Adjust term strengths to give best training set results then evaluate on test set
- E.g., training set: “My legs **ache**”: coder sentiment = 1,-3 => adjust sentiment of “ache” from -2 to -3.
Summary of sentiment methods

- **sentiment word strength list**
  - “miss” = +2, -2
  - **terrify** = -4

- **spelling corrected**
  - **nicce** -> nice

- **booster words** alter strength
  - **very** happy

- **negating words** flip emotions
  - **not** nice

- **repeated letters** boost sentiment/+ve
  - **niiiice**

- **emoticon list**
  - :) = +2

- **exclamation marks** count as +2 unless –ve
  - hi!

- **repeated punctuation** boosts sentiment
  - good!!!

- **negative emotion ignored in questions**
  - u h8 me?
Experiments

- Development data = 2600 MySpace comments coded by 1 coder
- Test data = 1041 MySpace comments coded by 3 independent coders
- Comparison against a range of standard machine learning algorithms
Test data: Inter-coder agreement

Krippendorff’s inter-coder weighted alpha = 0.5743 for positive and 0.5634 for negative sentiment

Only moderate agreement between coders but it is a hard 5-category task

<table>
<thead>
<tr>
<th>Comparison for 1041 MySpace texts</th>
<th>+ve agreement</th>
<th>-ve agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coder 1 vs. 2</td>
<td>51.0%</td>
<td>67.3%</td>
</tr>
<tr>
<td>Coder 1 vs. 3</td>
<td>55.7%</td>
<td>76.3%</td>
</tr>
<tr>
<td>Coder 2 vs. 3</td>
<td>61.4%</td>
<td>68.2%</td>
</tr>
</tbody>
</table>

Krippendorff's inter-coder weighted alpha = 0.5743 for positive and 0.5634 for negative sentiment.
# Results: +ve sentiment strength

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Optimal #features</th>
<th>Accuracy</th>
<th>Accuracy +/- 1 class</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiStrength</td>
<td>-</td>
<td>60.6%</td>
<td>96.9%</td>
<td>.599</td>
</tr>
<tr>
<td>Simple logistic regression</td>
<td>700</td>
<td>58.5%</td>
<td>96.1%</td>
<td>.557</td>
</tr>
<tr>
<td>SVM (SMO)</td>
<td>800</td>
<td>57.6%</td>
<td>95.4%</td>
<td>.538</td>
</tr>
<tr>
<td>J48 classification tree</td>
<td>700</td>
<td>55.2%</td>
<td>95.9%</td>
<td>.548</td>
</tr>
<tr>
<td>JRip rule-based classifier</td>
<td>700</td>
<td>54.3%</td>
<td>96.4%</td>
<td>.476</td>
</tr>
<tr>
<td>SVM regression (SMO)</td>
<td>100</td>
<td>54.1%</td>
<td>97.3%</td>
<td>.469</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>100</td>
<td>53.3%</td>
<td>97.5%</td>
<td>.464</td>
</tr>
<tr>
<td>Decision table</td>
<td>200</td>
<td>53.3%</td>
<td>96.7%</td>
<td>.431</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>100</td>
<td>50.0%</td>
<td>94.1%</td>
<td>.422</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>100</td>
<td>49.1%</td>
<td>91.4%</td>
<td>.567</td>
</tr>
<tr>
<td>Baseline</td>
<td>-</td>
<td>47.3%</td>
<td>94.0%</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>-</td>
<td>19.8%</td>
<td>56.9%</td>
<td>.016</td>
</tr>
</tbody>
</table>
SentiStrength vs. 693 other algorithms/variations

Results: -ve sentiment strength

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Optimal #features</th>
<th>Accuracy</th>
<th>Accuracy +/- 1 class</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (SMO)</td>
<td>100</td>
<td>73.5%</td>
<td>92.7%</td>
<td>.421</td>
</tr>
<tr>
<td>SVM regression (SMO)</td>
<td>300</td>
<td>73.2%</td>
<td>91.9%</td>
<td>.363</td>
</tr>
<tr>
<td>Simple logistic regression</td>
<td>800</td>
<td>72.9%</td>
<td>92.2%</td>
<td>.364</td>
</tr>
<tr>
<td><strong>SentiStrength</strong></td>
<td>-</td>
<td>72.8%</td>
<td>95.1%</td>
<td><strong>.564</strong></td>
</tr>
<tr>
<td>Decision table</td>
<td>100</td>
<td>72.7%</td>
<td>92.1%</td>
<td>.346</td>
</tr>
<tr>
<td>JRip rule-based classifier</td>
<td>500</td>
<td>72.2%</td>
<td>91.5%</td>
<td>.309</td>
</tr>
<tr>
<td>J48 classification tree</td>
<td>400</td>
<td>71.1%</td>
<td>91.6%</td>
<td>.235</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>100</td>
<td>70.1%</td>
<td>92.5%</td>
<td>.346</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>100</td>
<td>69.9%</td>
<td>90.6%</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>-</td>
<td>69.9%</td>
<td>90.6%</td>
<td>-</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>200</td>
<td>68.0%</td>
<td>89.8%</td>
<td>.311</td>
</tr>
<tr>
<td>Random</td>
<td>-</td>
<td>20.5%</td>
<td>46.0%</td>
<td>.010</td>
</tr>
</tbody>
</table>
Example differences/errors

THINK 4 THE ADD
- Computer (1,-1), Human (2,-1)

0MG 0MG 0MG 0MG 0MG 0MG 0MG 0MG
0MG!!!!!!!!!!!!!!!!!!!!!!!!N33N3R!!!!!!!!!!!!!!
- Computer (2,-1), Human (5,-1)
Application - Evidence of emotion homophily in MySpace

- Automatic analysis of sentiment in 2 million comments exchanged between MySpace friends
- Correlation of 0.227 for +ve emotion strength and 0.254 for –ve
- People tend to use similar but not identical levels of emotion to their friends in messages
SentiStrength 2

Sentiment analysis programs are typically domain-dependent

SentiStrength is designed to be quite generic
  - Does not pick up domain-specific non-sentiment terms, e.g., G3

SentiStrength 2.0 has extended negative sentiment dictionary
  - In response to weakness for negative sentiment

Thelwall, M., Buckley, K., Paltoglou, G. (submitted).
High Face Validity Sentiment Strength Detection for the Social Web
To test on a wide range of different Social Web text
SentiStrength 2 (unsupervised) tests

Social web sentiment analysis is less domain dependant than reviews

<table>
<thead>
<tr>
<th>Data set</th>
<th>Positive Correlation</th>
<th>Negative Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>0.589</td>
<td>0.521</td>
</tr>
<tr>
<td>MySpace</td>
<td>0.647</td>
<td>0.599</td>
</tr>
<tr>
<td>Twitter</td>
<td>0.541</td>
<td>0.499</td>
</tr>
<tr>
<td>Sports forum</td>
<td>0.567</td>
<td>0.541</td>
</tr>
<tr>
<td>Digg.com news</td>
<td>0.352</td>
<td>0.552</td>
</tr>
<tr>
<td>BBC forums</td>
<td>0.296</td>
<td>0.591</td>
</tr>
<tr>
<td>All 6</td>
<td>0.556</td>
<td>0.565</td>
</tr>
</tbody>
</table>
Why the bad results for BBC?

Long texts, mainly negative, expressive language used, e.g.,

- David Cameron must be very happy that I have lost my job.
- It is really interesting that David Cameron and most of his ministers are millionaires.
- Your argument is a joke.
SentiStrength vs. Machine learning for Social Web texts

- Machine learning performs a bit better overall (7 out of 12 data sets/+ve or negative)
  - Logistic Regression with trigrams, including punctuation and emoticons; 200 features
- But has “domain transfer” and “face validity” problems for *some* tasks
SentiStrength software

- Versions: Windows, Java, live online (sentistrength.wlv.ac.uk)
- German version (Hannes Pirker)
- Variants: Binary (positive/negative), trinary (positive/neutral/negative) and scale (-4 to +4)
- Sold commercially - & purchasers converting to French, Spanish, Portuguese, & ?
CYBEREMOTIONS = data gathering + complex systems methods + ICT outputs
Application – sentiment in Twitter events

- Analysis of a corpus of 1 month of English Twitter posts
- Automatic detection of spikes (events)
- Sentiment strength classification of all posts
- Assessment of whether sentiment strength increases during important events

Automatically-identified Twitter spikes

Proportion of tweets mentioning keyword
Chile

matching posts

Proportion of tweets mentioning Chile

Date and time

Increase in -ve sentiment strength

Av. +ve sentiment
Just subj.

Av. -ve sentiment
Just subj.
#oscars

Proportion of tweets mentioning the Oscars

Increase in $-ve$ sentiment strength

Av. $+ve$ sentiment
Just subj.

Av. $-ve$ sentiment
Just subj.
Sentiment and spikes

Analysis of top 30 spiking events

- Strong evidence (p=0.001) that higher volume hours have stronger negative sentiment than lower volume hours

- Insufficient evidence (p=0.014) that higher volume hours have different positive sentiment strength than lower volume hours

=> Spikes are typified by increases in negativity
But there is plenty of positivity if you know where to look!
YouTube Video comments

- Short text messages left for a video by viewers
- Up to 1000 per video accessible via the YouTube API
- A good source of social web text data
Sentiment in YouTube comments

Predominantly positive comments
Trends in YouTube comment sentiment

- +ve and –ve sentiment strengths negatively correlated for videos (Spearman’s rho -0.213)
- # comments on a video correlates with –ve sentiment strength (Spearman’s rho 0.242, p=0.000) and negatively correlates with +ve sentiment strength (Spearman’s rho -0.113) – negativity drives commenting even though it is rare!
- Qualitative: Big debates over religion
  - No discussion about aging rock stars!
Conclusion

- Automatic classification of sentiment strength is possible for the social web – even unsupervised!
  - ...not good for longer, political messages?
- Hard to get accuracy much over 60%?
- Can identify trends through automatic analysis of sentiment in millions of social web messages