Evaluating Sentiment Analysis Evaluation

Gerald Penn, University of Toronto
http://www.cs.toronto.edu/~gpenn
5th March, 2014
Typical Evaluation of a Sentiment Analyzer

1) Build the sentiment analyzer
   – This normally involves estimating statistical parameters of a model using labelled training data

2) Classify the sentiment of some “held-out” data (*development test set*)
   – The held-out data are often sampled from the same source as the training data

3) Compute classifier accuracy scores
   – Precision, Recall, F-measure, Accuracy

4) If not good enough, react by improving the method and goto (2)

5) Evaluate your sentiment analyzer on some new held-out data (*evaluation test set*)

6) Again, compute classifier accuracy scores

This protocol was developed by engineers – it is the discrete analogue to a Receiver Operating Characteristics (ROC) curve.
ROC-style Evaluations are Different

• ROC-style evaluations were created in an environment where:
  – Making prototypes is labour-intensive and costly
  – Sampling data for training, development testing or evaluation is (comparatively) cheap

• But in sentiment analysis (and a number of other technologies):
  – Tweaking prototypes is not nearly as expensive as…
  – …collecting labelled data, which generally involves recruiting and incentivizing human annotators/judges.

• In sentiment analysis, we’re also dealing with essentially human sentiments.
The Usual Answer: lower the cost of annotation

- Mechanical Turk
- “Gamification”
- Different incentivization schemes
- Even sentiment analysis itself, although to builders of sentiment analyzers, that isn’t much help.
How are Evaluations conducted in the Social Sciences?

• Includes economics, human-computer interaction (HCI) etc.

• There are always two components to an evaluation:
  – Subjective: ask people what they think.
  – Objective: give them a task to perform and watch them vote with their feet.

• It is almost unheard of in academic research to gather truly objective labelled data.
  – It’s too expensive.
  – Experiments must be very carefully controlled to preserve “ecological validity.”
How are Evaluations conducted in the Social Sciences?

• Instead, what we usually collect is a lot of this:

Asking human judges to label sentiment without a proper task-embedding

• …and we even have the nerve to call it objective or ground truth data – after all, it’s a real human.
Example: Film Reviews

• Ask this guy to rate your films and write about them

• …or worse still, ask him to rate a review someone else has written, without having watched the film himself

• … but don’t look at box-office returns, Academy awards, exit interviews with viewers, etc.
Example: Consumer Product Advertising

• Ask this guy to watch your commercial or use your product and then express his sentiment about the product

• …but don’t go shopping with him to see what he buys.
Example: Stock Trading

• Ask this guy to label quarterly stock reports/news articles, blogs, tweets about a publicly traded company

• …but don’t look at how the market itself reacts/has reacted.
Awww, do we really have to?

• I regret to inform you that the difference actually does matter.
• There is solid empirical evidence to show that, if you’re building a sentiment analyzer to predict the market movements of equities (for example), you’d better go out and measure how the market moves.
• But there is some good news: it even helps to:
  – train on pseudo-objective, human-annotated data
  – …but then tune using “development” test data that have been properly collected.
• Here’s an example of this in the context of sentiment-based market-neutral equity trading…
How to make a pile of money off sentiment analysis

• SVM classifier with a linear kernel
• Trained on linguistic features extracted from Reuters news documents on the topic of NYSE-listed companies
  – very simple features: word frequencies weighted by the BM25 scheme (Paltoglou and Thelwall, 2010), excluding a stop list.
• We randomly sampled a list of NYSE-traded companies as at March, 1997, balanced over three levels of market capitalization (small, mid, large).
• Then we collected every third Reuters document about those companies.
How to make a pile of money off sentiment analysis (2)

• Our test data consisted of those reports published during or after March, 2005, which we further subdivided into development and evaluation test data.

• Our training data consisted of those reports published between March, 1997 and March, 2005, mixed with a separate collection of documents sampled from Reuters again on companies that were still being trained as at March, 2013 (survivor bias).

• There were 1,256 documents in total.

• These were labelled by two judges as positive (+1), neutral (0) or negative (-1).
Classifier Accuracy

• With our features: 79.827% accuracy.
• With normalized 0/1 word-presentation features: 80.164%
  – Pang and Lee (2004) got 86.4% on film reviews. We got 86.85%.
• In the financial domain, many others report around 70% (Koppel and Shtrsimberg, 2004).
• So this is reasonable.
• But now let’s throw away the test set labels and instead go make some trades.
Task-embedded Evaluation

- Recall that our data were labelled positive, neutral or negative.
- So we buy, hold or sell accordingly, for some period of time.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Period</th>
<th>Return</th>
<th>S. Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>30 days</td>
<td>-0.037%</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>0.763%</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>0.742%</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.716%</td>
<td>0.108</td>
</tr>
<tr>
<td>Momentum</td>
<td>30 days</td>
<td>1.176%</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>0.366%</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>0.713%</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.017%</td>
<td>-0.002</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>30 days</td>
<td>0.318%</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>-0.038%</td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>-0.035%</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.046%</td>
<td>0.036</td>
</tr>
<tr>
<td>Oracle S&amp;P</td>
<td>30 days</td>
<td>3.765%</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>1.617%</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>1.390%</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.860%</td>
<td>0.909</td>
</tr>
<tr>
<td>Oracle</td>
<td>30 days</td>
<td>11.680%</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>5.143%</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>4.524%</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>3.542%</td>
<td>0.630</td>
</tr>
</tbody>
</table>
Tuning with market data

• But we can do better – instead of using the sentiment class, we can use the underlying sentiment score provided by the SVM.

• This is now a regression problem: if the score is above threshold \( p \), we’ll go long, if it’s below \( n < p \), we’ll go short.

• Zhang and Skiena (2010) did something similar in an equity trading strategy, with great improvements to their returns.

• But we’ll determine these parameters empirically by trading for a bit to determine the consequences of different settings.

• The result? Better – about double the return (and the 30-day trader makes money).
Tweaking with market data

• But I want more, MORE! So let’s try to change the linguistic features that we trade on – maybe there’s some improvement to be gained from this.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Accuracy</th>
<th>30 days</th>
<th>5 days</th>
<th>3 days</th>
<th>1 day</th>
</tr>
</thead>
<tbody>
<tr>
<td>term_presence</td>
<td>80.164%</td>
<td>3.843%</td>
<td>1.851%</td>
<td>1.691%</td>
<td>2.251%</td>
</tr>
<tr>
<td>bm25_freq</td>
<td>81.143%</td>
<td>1.110%</td>
<td>1.770%</td>
<td>1.781%</td>
<td>0.814%</td>
</tr>
<tr>
<td>bm25_freq_d_n_copular</td>
<td>62.094%</td>
<td>3.458%</td>
<td>2.834%</td>
<td>2.813%</td>
<td>2.586%</td>
</tr>
<tr>
<td>bm25_freq_with_sw</td>
<td>79.827%</td>
<td>0.390%</td>
<td>1.685%</td>
<td>1.581%</td>
<td>1.250%</td>
</tr>
<tr>
<td>freq</td>
<td>79.276%</td>
<td>1.596%</td>
<td>1.221%</td>
<td>1.344%</td>
<td>1.330</td>
</tr>
<tr>
<td>freq_with_sw</td>
<td>75.564%</td>
<td>1.752%</td>
<td>0.638%</td>
<td>1.056%</td>
<td>2.205%</td>
</tr>
</tbody>
</table>

• Which would you believe: classifier accuracy or trading?
• Using trading, we got as high as 70% annual return.
Conclusion

• Collect properly controlled task-embedded data – it really isn’t clear what the other stuff is telling you.

• Even if you only collect a little, your results can get much better, just through some tuning.

• Sentiment scoring is as valuable as sentiment classification.
Thanks! Questions?

Gerald Penn, University of Toronto
http://www.cs.toronto.edu/~gpenn
5th March, 2014