

Compositional Sentiment Analysis

Stephen Pulman

Dept. of Computer Science, Oxford University
and TheySay Ltd

`stephen.pulman@cs.ox.ac.uk`, `stephen.pulman@theysay.io`

June 3rd 2014



TheySay

What is sentiment analysis?

The term has been around since 2000ish, and has been used to cover a variety of different phenomena:

Sentiment proper

Positive, negative, or neutral attitudes expressed in text:

Suffice to say, Skyfall is one of the best Bonds in the 50-year history of moviedom's most successful franchise.

Skyfall abounds with bum notes and unfortunate compromises.

There is a breach of MI6. 007 has to catch the rogue agent.

Sentiment analysis = emotion detection?

Emotion

A variety of different theories of emotional state:

- Ekman: anger, disgust, fear, happiness, sadness and surprise.
- multi-dimensional theories:
 - ▶ Pleasure vs. displeasure (how pleasant?)
 - ▶ Arousal vs. non-arousal (how intense)
 - ▶ Dominance vs. submissiveness (e.g. anger vs. fear)
- psychometric measures 'calmness', 'vitality', etc....
- activation, valence, potency, emotion intensity

Typically it is assumed that particular words and phrases are associated with these categories.

Emotion classification is difficult:

Emotion: human annotation usually gives upper limit on what is possible. For Ekman emotion labels the results are not good: this table shows level of agreement between annotators on 1250 news headlines drawn from major newspapers such as New York Times, CNN, and BBC News, as well as from the Google News search engine. Performance of various different types of classifiers on a collection of annotated blog posts is shown in the third column.

Emotions	human agreement	best classifier F score
anger	49.55	16.77
disgust	44.51	4.68
fear	63.81	22.8
joy	59.91	32.87
sadness	68.19	23.06
surprise	36.07	14.1

Sentiment analysis = speculation, intent, etc?

Modality and speculation

Particularly in clinical or scientific texts:

Certain: This demonstrates/proves/disproves that...

Possible: This suggests that..., these results are consistent with...

Also a variety of '**hedges**': almost, nearly, broadly

And some very domain dependent phenomena: "Earnings were broadly in line with expectations"

Risk, future intent detection

Usually in a very specific domain:

- detection of future predictions or commitments in financial reports

- 'intent to churn' signals in blogs or CRM messages:

"Terrible service... Paypal should take responsibility for accounts which have been hacked into ... **Very disappointed and will never use Paypal again.**"

Building a sentiment analysis system

Version 1: cheap and cheerful

- collect lists of positive and negative words or phrases, from public domain lists or by mining them.
- given a text, count up the number of positives and negatives, and classify based on that.
- you would be surprised how many commercial systems seem to do no more than this.

Problems:

- if number of positive = number of negatives, do we say 'neutral'?
- **Compositional sentiment:** a phrase like 'not wonderfully interesting' is negative, even though 'wonderfully' and 'interesting' will be in the list of positive words.
- some words positive in some contexts, negative in others: 'cold beer' is good, 'cold coffee' is not. (This is actually a problem for all approaches.)

Version 2: better (what most commercial systems do)

A bag-of-words classifier:

- get a training corpus of texts human annotated for sentiment (e.g. **pos**/**neg**/**neut**).
- represent each text as a vector of counts of n -grams¹ of (normalised) words, and train your favourite classifier on these vectors.
- should capture some ‘compositional’ effects: e.g. ‘very_interesting’ likely signal for positivity, whereas ‘not_very’ a signal for negativity.
- will work for any language and domain where you can get accurately labelled training data.
- bag-of-words means structure is ignored:
“Knox is found guilty of the murder of Kercher”
= “Kercher is found guilty of the murder of Knox”

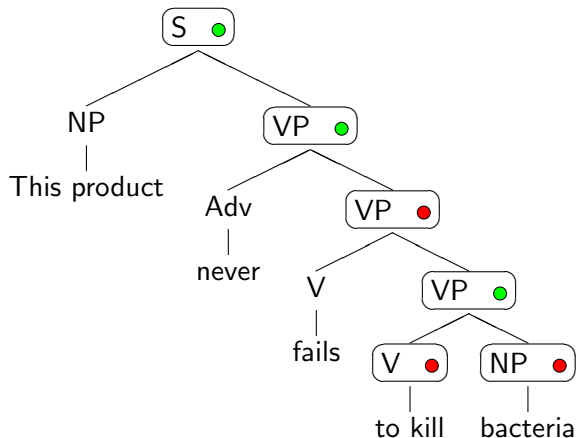
¹ n usually ≤ 3 , and as n gets bigger, more training data is required

Problems:

- Equally balanced texts will still be problematic,
- and richer compositional effects will still be missed:
 - clever, too clever, not too clever
 - bacteria
 - kill bacteria
 - fail to kill bacteria
 - never fail to kill bacteria
- difficult to give sentiment labels accurately to short units like sentences or phrases,
- or to pick out mixed sentiment:
 - “The display is amazingly sharp. However, the battery life is disappointing.”
- Complex compositional examples occur quite frequently in practice:
 - The Trout Hotel: This newly refurbished hotel could not fail to impress...
 - BT: it would not be possible to find a worse company to deal with...

Version 3: best - use linguistic analysis

- do as full a parse as possible on input texts.
- use the syntax to do 'compositional' sentiment analysis:



Sentiment logic rules²


- kill + negative → positive (kill bacteria)
- kill + positive → negative (kill kittens)
- too + anything → negative (too clever, too red, too cheap)
- etc. In our system (www.thesay.io) we have 65,000+ of such rules...

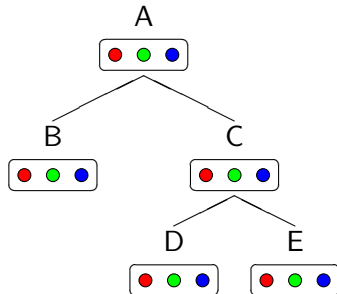
Problems:



- still need extra work for context-dependence ('cold', 'wicked', 'sick'...)
- can't deal with reader perspective: "Oil prices are down" is good for me, not for Chevron or Shell investors.
- can't deal with sarcasm or irony: "Oh, great, they want it to run on Windows"

²Moilanen and Pulman, 2007

Machine learning for composition³

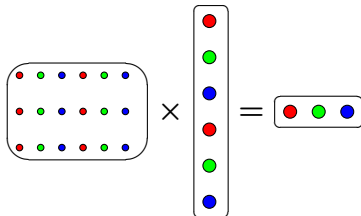
Assume we have a 'sentiment treebank'. Represent words as vectors .



To compute C's vector, we concatenate those of D and E , and learn from the training data a function which combines them in the 'right' way to form another . Likewise we combine B and C to find A.

³Hermann and Blunsom, 2013; Socher et al 2013

Various composition functions

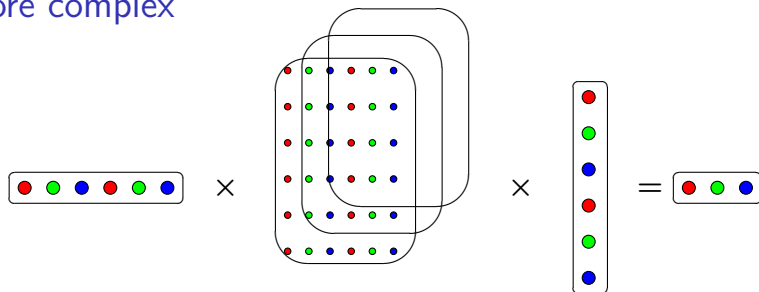


Here the weight matrix represents pairwise word/phrase combinations, perhaps also with syntactic info. We apply an element-wise non-linearity to the resulting vector. Weights can be learned via neural network methods.

We then use a 'softmax' function to map from the phrase vectors to a distribution over sentiment labels:

$$\begin{bmatrix} \text{red} & \text{green} & \text{blue} \end{bmatrix} \text{ Softmax} \Rightarrow \{ \text{Pos} = 0.5, \text{Neg} = 0.3, \text{Neut} = 0.2 \}$$

More complex



Here a bilinear tensor operation combines the concatenated child vectors with their transpose: each slice of the tensor represents a different composition operation. We can combine this with the previous compositional function.

Socher's system, trained and tested on a treebank derived from a standard movie review corpus gives better results on both multi-label and binary sentiment analysis than non-compositional systems.

Some applications

The most obvious and widely used are in brand and reputation management, typically along these lines:

- release a new product
- track consumer reaction
- find out what are the most/least popular features
- get comparisons with rival products

On the consumer side, you could in principle use sentiment analysis to construct a personalised 'Which?' report comparing e.g. different cameras.

In some areas, it is possible to predict sales on the basis of sentiment analysis: e.g. Liu and his team claim to be able to predict box-office revenues for movies.

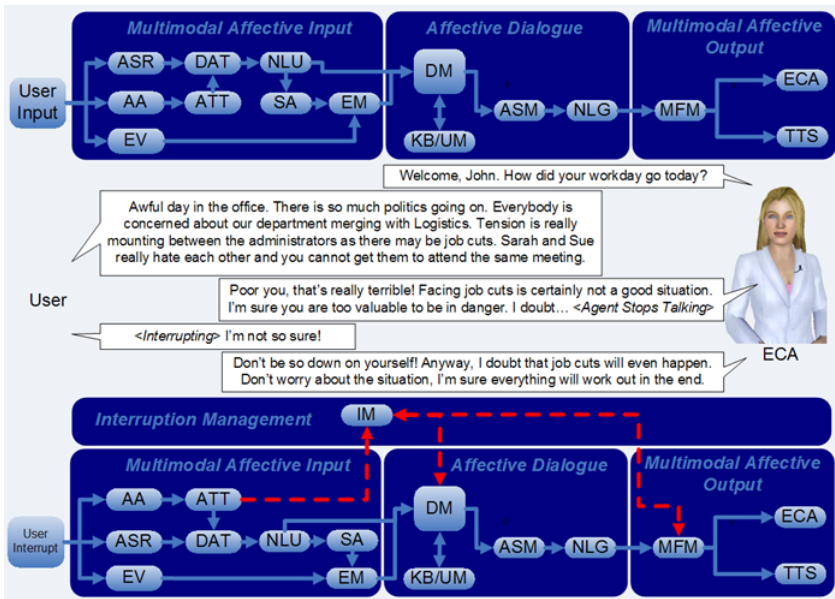
Note that there is a whole sub-industry in generating fake positive or negative reviews to skew results and damage your rivals...and a sub-sub-industry in trying to detect this kind of spam!

Sentiment analysis in dialogue

The Companions project (www.companions-project.org)

- €12m EU-funded project over 4 years with 14 partners (!)
- aimed to create a multi-modal 'companion'.
- speech recognition and synthesis, dialogue processing, avatar with gesture and facial expression.
- used text sentiment analysis and speech emotion detection (EmoVoice).
- several scenarios: most developed was the 'How was your day?' post-work chat.
- system aimed to detect mood and respond appropriately.
- system interruptible by user (not as easy as you might think...)

Conversation about the day...



Speech and text sentiment

- Process user input for sentiment, both speech and text.
- Use this to guide the generation of a response
- If very negative, e.g. 'You seem rather upset'
- Or very positive, 'Great! Glad to hear that.'

We can also invert these polarities to generate ironical or sarcastic replies:

User:

At my performance review I got offered a promotion but it just means more hard work and responsibility.

System:

Instead of:

Well done: everyone likes getting a promotion. It will further your career - that's good.

We get:

Who likes getting a promotion anyway? You will further your career, isn't that awful?

An experiment we didn't have time for

Can you **recognise** sarcasm and irony by exploiting mismatches between speech and text properties?

Positive text + negative speech:

What a wonderful start to the day!

Oh, that's just great!

Negative text + positive speech?

Do we get the opposite? Seems intuitively to be much less common:

Hey, yet another disaster!

Absolutely awful screw-up!

Betting on horse racing

Large online betting community, using Twitter, Betfair.com forums etc.



Hypothesis: profile opinion about particular horses and riders leading up to a race to detect mismatches between sentiment and bookmakers' odds.

Result: dismal failure, at least using Twitter! Could not separate signal from noise:

- Many horse names (up to 300 per meeting).
- Horse names not easily recognisable as such, e.g. 'Genuine Art', 'Beat That', 'Talent', 'Ask the Boss', 'Degenerate', and even 'Say'
- Many racing specific phrases: 'will not like the going', 'pulls up', 'might not stay' . . .
- Horse specific info: 'going is soft' - good for some, not others

Lesson learned: many data sources need sophisticated pre-processing to get accurate sentiment analyses. . .

Predicting Man of the Match

Method: track player Tweet mentions during game. Predict that player with highest proportion of positive tweets at end of game will be MOTM.

Results: prediction correct about 50% of the time. MOTM usually in top 3 players, but events near the end of a match often assume greater importance and affect result:



Predicting MOTM

Mixed success:

England vs. Scotland (friendly), 14th Aug 2013

Top ten most positive entities:

Rank	entity
1	#engsco
2	I
3	goal
4	Miller - highest ranking player
5	game
6	half
7	Scotland
8	England
9	Welbeck - actual MOTM
10	we

Not enough data points...

- ▶ Manchester United vs. Wigan Athletic, 10th Aug 2013
Top mentioned player (rank 11) Robin van Persie, who was voted MOTM
- ▶ Manchester United vs. Swansea, Aug 17th
Robin van Persie narrowly ahead of Rooney (rank 4), and van Persie again MOTM.
- ▶ Manchester City vs Newcastle 19 August 2013
Top ranked player Tim Krul (Newcastle, rank 8)
but Edin Dzeko was declared the MotM, from Manchester City.
- ▶ Barcelona vs Levante, 19 August, top ranked PERS entities
Neymar (3) and @piersmorgan (9) !!!
In fact Messi and Pedro were joint MOTM

In the Spanish game Neymar only joined the game from the bench in the 63rd minute, probably accounting for his tweet volume.

Predicting election results

Can we eliminate opinion polls?

A very mixed picture, to put it mildly:

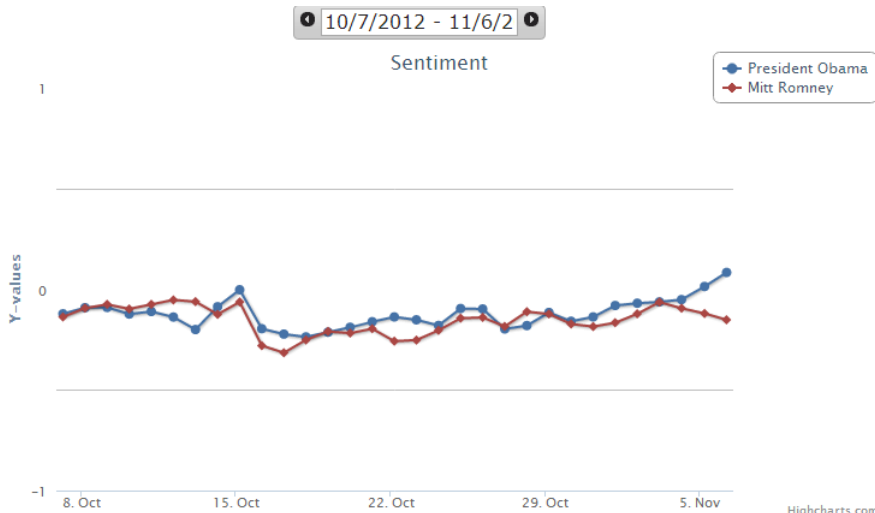
- Tumasjan et al. 2010 claimed that share of volume on Twitter corresponded to distribution of votes between 6 main parties in the 2009 German Federal election. (Volume rather than sentiment is also a better predictor of movie success).
- Jungherr et al. 2011 pointed out that not all parties running had been included and that different results are got with different time windows. Tumasjan et al. replied, toning down their original claims.
- Tjong et al. found that Tweet volume was NOT a good predictor for the Dutch 2011 Senate elections, and that sentiment analysis was better.

Predicting election results

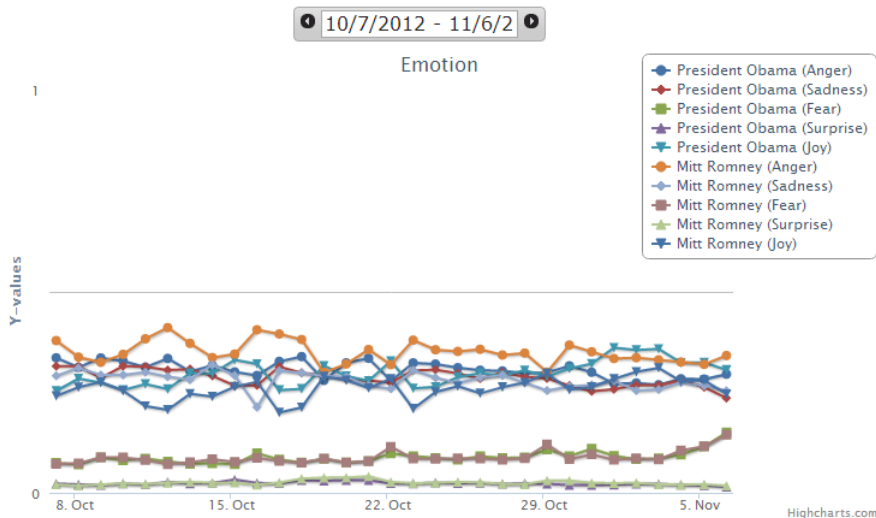
- Skoric et al. 2012 found some correlation between Twitter volume and the votes in the 2011 Singapore general election, but not enough to make accurate predictions.
- Bermingham and Smeaton 2011 found that share of Tweet volume and proportion of positive Tweets correlated best with the outcome of the Irish General Election of 2011
- ... but that the mean average error was greater than that of traditional opinion polls!

So this doesn't look very promising, although of course other sources than Twitter might give better results - but they are difficult to get at quickly.

Sentiment Obama vs. Romney (from twitris.knoesis.org)



Emotion Obama vs. Romney



Financial market prediction

Looking for 'alpha' signal: Bollen et al 2011.

- harvest tweets from 2008 containing 'mood' indicators ('I feel.. I am feeling.. ... makes me...' etc)
- process tweets using OpinionFinder (positive/negative) and GPOMS (profile of mood states)
- GPOMS measures mood: Calm, Alert, Sure, Vital, Kind and Happy by looking for c.950 words or phrases that are correlated with these dimensions. This tool is derived from 'an existing and well-vetted psychometric instrument'...

Over a period of a few weeks which included the US 2008 election and Thanksgiving, found correlation between these events and positivity, + GPOMS dimensions Calm, Sure, Happy, and Vital...

Some correlations if you look hard enough...

They looked at correlations between these signals and the Dow Jones Industrial Average, fitting a regression model:

$$D_t = \alpha + \sum_{i=1}^n \beta_i * D_{t-i} + \sum_{i=1}^n \gamma_i * X_{t-i} + \epsilon_t$$

DJIA values represented as differences between day d and d-1. The β_i values range from 1 to 3 representing lags. The X ranges over the different sentiment and GPOMS dimensions. All quantities are normalised to standard deviations around the mean.

They found that the only significant correlation was at lag 3 for the dimension 'Calm'...

But can you make money?

Financial News, 15 Aug 2011

“A hedge fund that uses Twitter data to drive its trading strategy returned 1.85% in its first month of trading, according to an investor in the fund, in the first sign that social media data can be used successfully to enhance electronic trading techniques.

Derwent Capital, which finished its first month of trading at the end of July, beat the S&P 500 which fell 2.2% in July, while the average hedge fund made 0.76%, according to Hedge Fund Research...”

Financial Times, May 24, 2012

Last tweet for Derwent's Absolute Return

By James Mackintosh, Investment Editor

“The only dedicated “Twitter” hedge fund has shut down after deciding to offer its social media indicators to day traders instead. Derwent Capital Markets Absolute Return fund was quietly liquidated just a month after starting up last year...”

Predicting the US non-farm payroll⁴

NFP: a monthly economic index that measures job growth or decay:

- a 'market mover'.



Questions:

- Can we predict the direction of the NFP from financial indicators?
- Can we predict the direction of the NFP from sentiment in text?
- If so, does compositional sentiment perform better than BOW classifier?

⁴Joint work with Oxford Man Institute of Quantitative Finance

Back tested over data from 2000-2012

Almost 10m words of text containing relevant keys:

Source	Sentences
Associated Press	54K
Dow Jones	236K
Reuters	169K
Market News	385K
Wall Street Journal	76K

- and financial time-series data from many different sources, including:

Consumer Price Index (CPI)

Institute of Supply Management manufacturing index (ISM)

Job Openings and Labor Turnover Survey (JOLTS)

Process text using TheySay's API:

"The Governor noted that despite jobs being down, there was a surprising bright spot: construction added 1,900 jobs in November - its largest gain in 22 months."

pos: 0.925, neg: 0.0, neut: 0.075, conf: 0.69

"When I drive down the main street of my little Kansas City suburb I see several dark empty storefronts that didn't used to be that way."

pos: 0.0, neg: 0.973, neut: 0.027, conf: 0.674

"We continue to fare better than the nation - our rate has been at or below the national rate for 82 out of the past 83 months - but we must also recognize that there were 10200 jobs lost at the same time."

pos: 0.372, neg: 0.591, neut: 0.037, conf: 0.723

Test bag of words vs. compositional sentiment

Method:

- Also train a 1-3gram Support Vector Machine BOW sentiment classifier for comparison.
- Train individual logistic regression classifiers on each text and numerical stream.
- Use output sentiment distributions per time-slice as input feature for LR classifiers.
- Use a novel 'Independent Bayesian Classifier Combination' method to get best combination of individual classifiers.

BOW does not help classifier combination (AUC scores)

Individual classifiers for each stream:		A	B	C
	AP	0.59	0.69	0.37
	Dow Jones	0.45	0.44	0.25
	Reuters	0.50	0.46	0.36
	Market News	0.66	0.70	0.23
	Other Sources	0.58	0.63	0.63
	WSJ	0.44	0.63	0.53
Combined classifiers:		0.67	0.81	0.85

A = **average** % of **pos/neg** per time slice from SVM classifier trained on 1-3gram features

B = **average** % of **pos/neg/neut** per time slice from compositional sentiment as features

C = **trends**, i.e. differences between B compositional averages in successive time slices as features.

BOW info does not enable the IBCC method to improve on best individual classifier, but the compositional method does, by a huge margin.

Text combined with numerical streams

Numerical time series data alone does a good job:

Source	AUC
CPI	0.70
ISM	0.85
JOLTS	0.66
LFL	0.71
Combined	0.90

But a combination of text and time series is best!

Source	AUC
Time Series + Text Averages	0.94
Time Series + Text Trends	0.91

Conclusions

- Compositional sentiment methods give a substantial improvement in accuracy
- ...and finer grained analyses
- ...and in the financial domain at least can yield accurate predictions, especially when combined with numerical data.

Try it out!

www.theysay.io

<http://nlp.stanford.edu:8080/sentiment/rntnDemo.html>

References

Karo Moilanen and Stephen Pulman, 2007, *Sentiment composition*. In Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007), pages 378382, Borovets, Bulgaria, September 27-29 2007.

http://www.clg.ox.ac.uk/_media/people:karo:sentcompranlp07final.pdf

Karl Moritz Hermann and Phil Blunsom, 2013, *The Role of Syntax in Vector Space Models of Compositional Semantics*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 894904. Sofia, Bulgaria.

<http://aclweb.org/anthology//P/P13/P13-1088.pdf>

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng and Chris Potts, 2013, *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank* Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, Washington, USA; ACL, pp. 1631-1642.

<http://aclweb.org/anthology//D/D13/D13-1170.pdf>

Abby Levenberg, Stephen Pulman, Karo Moilanen, Edwin Simpson and Stephen Roberts, 2014, *Predicting Economic Indicators from Web Text Using Sentiment Composition*, International Journal of Computer and Communication Engineering. Barcelona, Spain. February, 2014. IACSIT Press.

http://www.oxford-man.ox.ac.uk/sites/default/files/sentiment_ICICA2014.pdf