Lexical resources

Text Analytics - Andrea Esuli

The language of opinions

The language we use to express our subjective evaluations is one of the most complex parts of language.

There are many components in the language of opinions:

- Global/Domain-specific lexicon.
- Valence shifters/Comparative expressions.
- Irony, sarcasm, common knowledge.
- . . .

A list of sentiment-relevant words can be the starting point to recognize and model new features extracted from a piece of text to determine its sentiment.

The language of opinions

Some **words** have a **globally** recognized **sentiment valence** in any context of use, e.g.: *"good", "poor", "perfect", "ugly"*

"A good tool that works perfectly"

"I had an horrible experience"

The simple **presence** of one of this words in text can be strong clue about the sentiment expressed in text.

Global lexicons

General purpose lexical resources list these words associating sentiment labels to them, e.g.:

- The General Inquirer lexicon
- MPQA
- WordNet affect
- SentiWordNet
- Appraisal lexicon

Global lexicons can be used to model **new features** that are extracted from text or as starting information to create a domain specific lexicon.

General Inquirer

The General Inquirer is a text analysis tools developed in the ’60.

It used a combination of a number of lexicons that label 11,788 words with respect to 182 semantic categories, including both topic and sentiment-related concepts.

The *Positiv* and *Negativ* categories are the largest ones, comprising 4,306 words.

ABILITY	*Positiv*	*Strong*	*Virtue*	*Eval*	*Abstract*	*Means*	*Noun*
ACCOMPLISH	*Positiv*	*Strong*	*Power*	*Active*	*Complet*	*Verb*	

MPQA lexicons

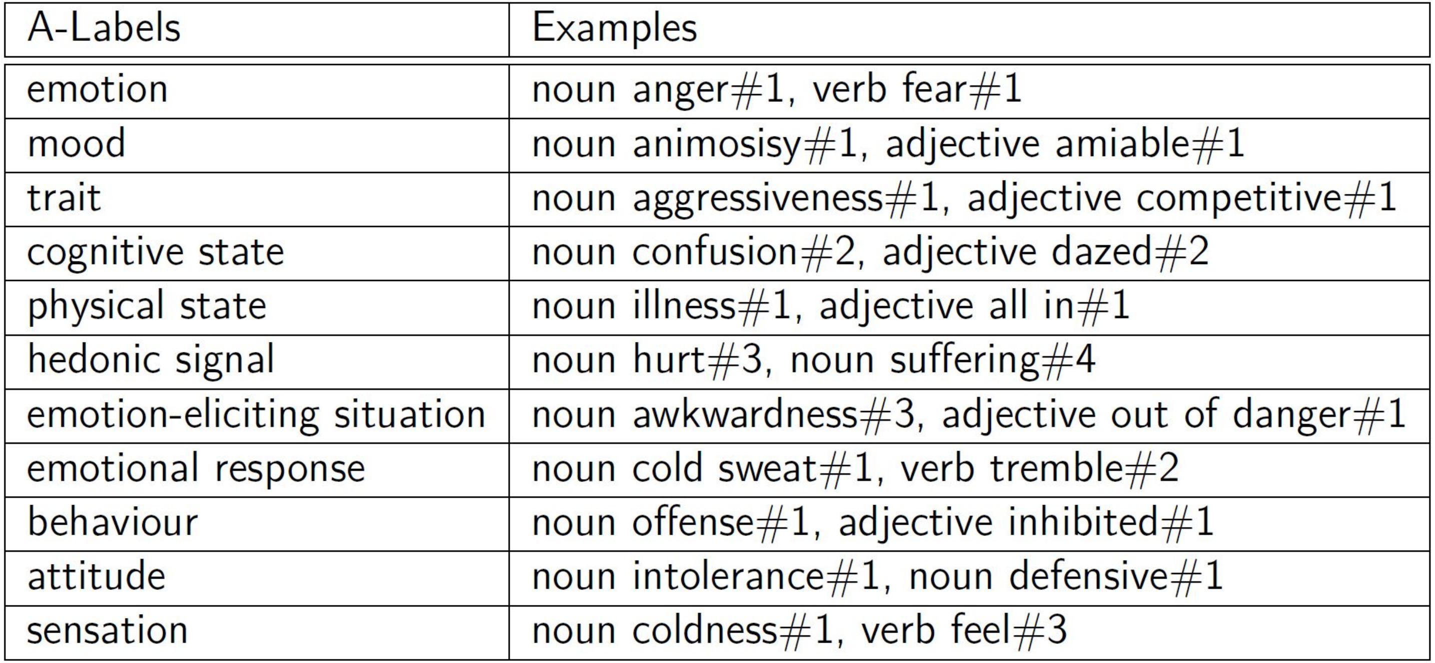
The Multi Perspective Question Answering project produced a number of lexicons for sentiment-related tasks:

- Subjectivity Lexicon: a list 8k+ words that are clues for subjectivity
- Subjectivity Sense Annotations: word senses with subjectivity labels
- Arguing Lexicon: patterns related to arguing, classified w.r.t. different types of arguments.
- +/-Effect Lexicon: 880 hand-labeled + 11k automatic-labeled word senses about the effect they have on the event they are related to, e.g.:

The bill would curb skyrocketing health care costs [source]

WordNet affect

WordNet affect annotates WordNet synsets with emotion-related labels.



A-Labels	Examples
emotion	noun anger#1, verb fear#1
mood	noun animosisy#1, adjective amiable#1
trait	noun aggressiveness#1, adjective competitive#1
cognitive state	noun confusion#2, adjective dazed#2
physical state	noun illness#1, adjective all in#1
hedonic signal	noun hurt#3, noun suffering#4
emotion-eliciting situation	noun awkwardness#3, adjective out of danger#1
emotional response	noun cold sweat#1, verb tremble#2
behaviour	noun offense#1, adjective inhibited#1
attitude	noun intolerance#1, noun defensive#1
sensation	noun coldness#1, verb feel#3

SentiWordNet

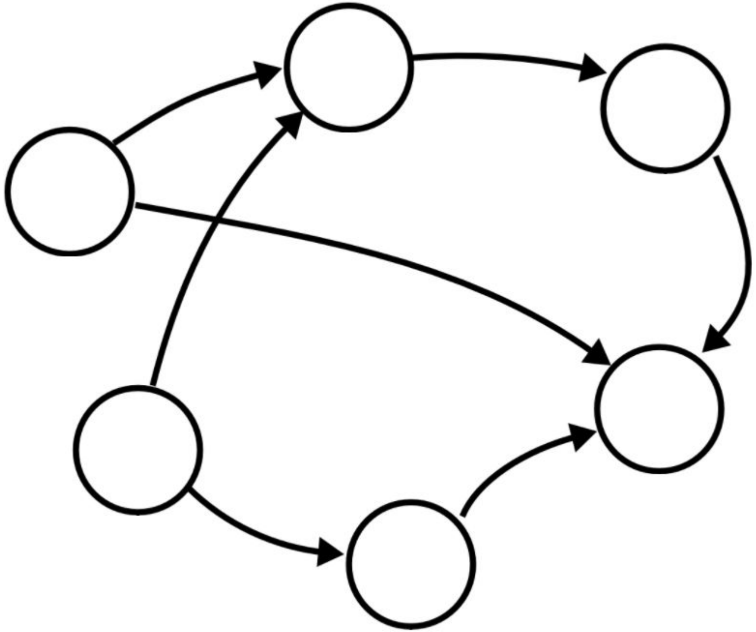
SentiWordNet assigns to each synset of WordNet a triple of sentiment scores: positivity, negativity, objectivity.

SentiWordNet is generated by applying a **random walk** process to the graph of WordNet synsets.

- Each synset is a node.
- A link between a and b exists if a appears in the gloss of b.
- (Sentiment) properties flow through links.

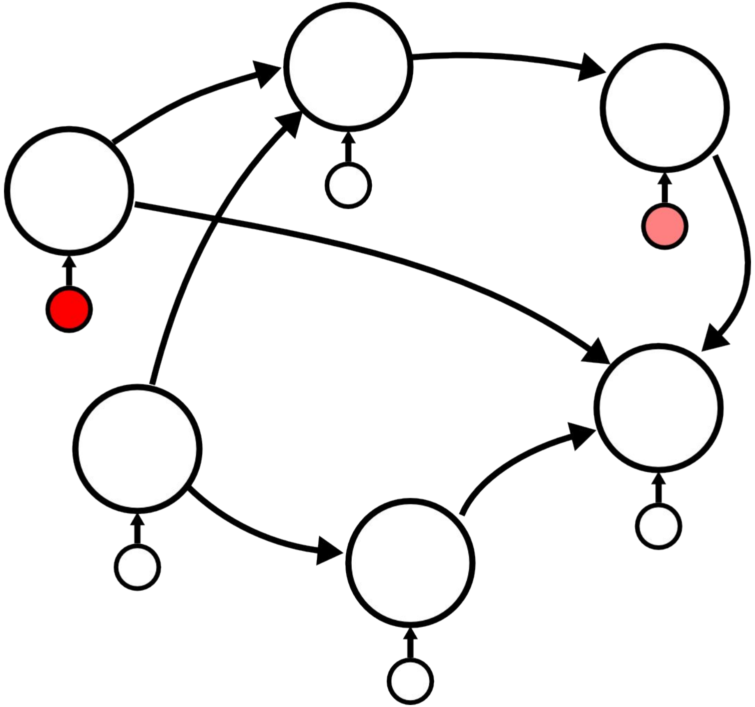
SentiWordNet

The graph is built by analyzing the glosses in WordNet.



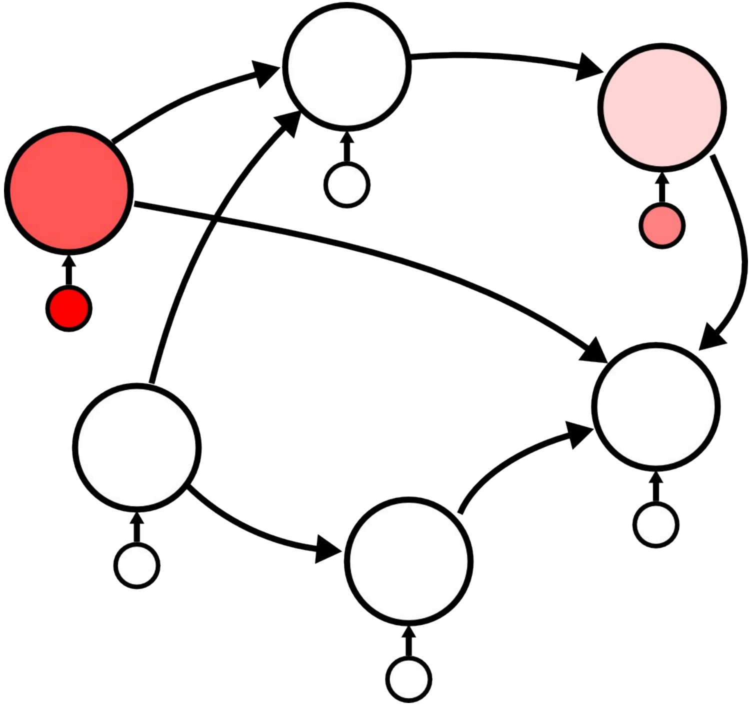
SentiWordNet

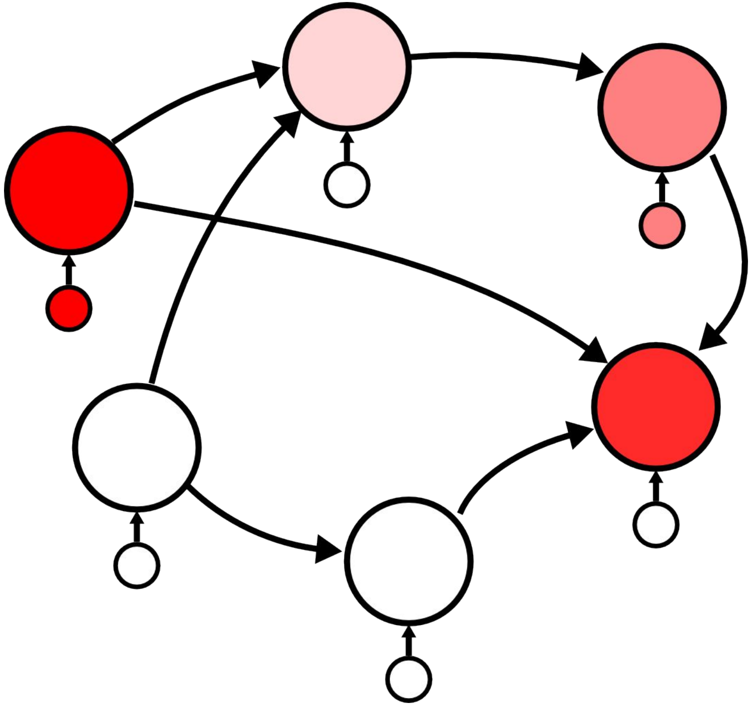
A few words with strong polarity are marked as sources of sentiment.



SentiWordNet

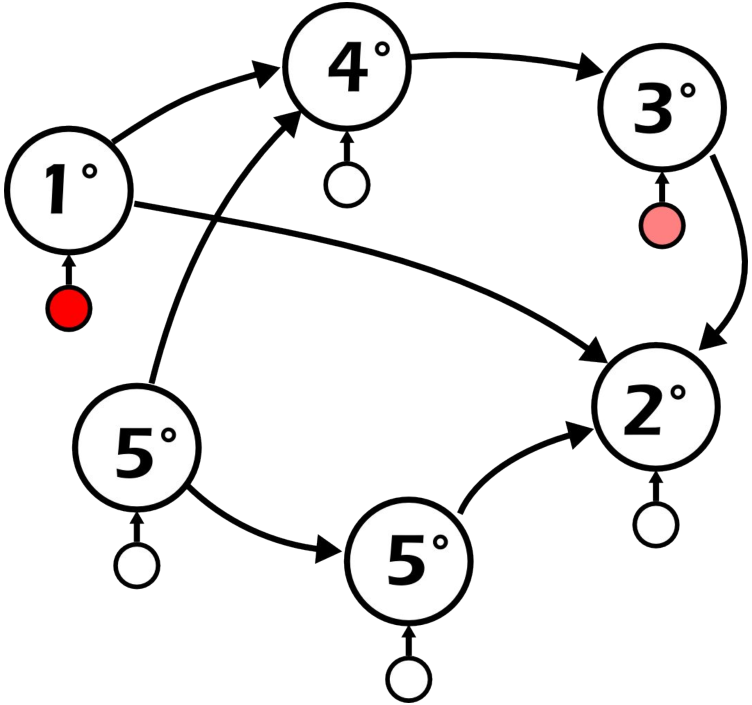
Pagerank is applied to the graph to spread sentiment values.



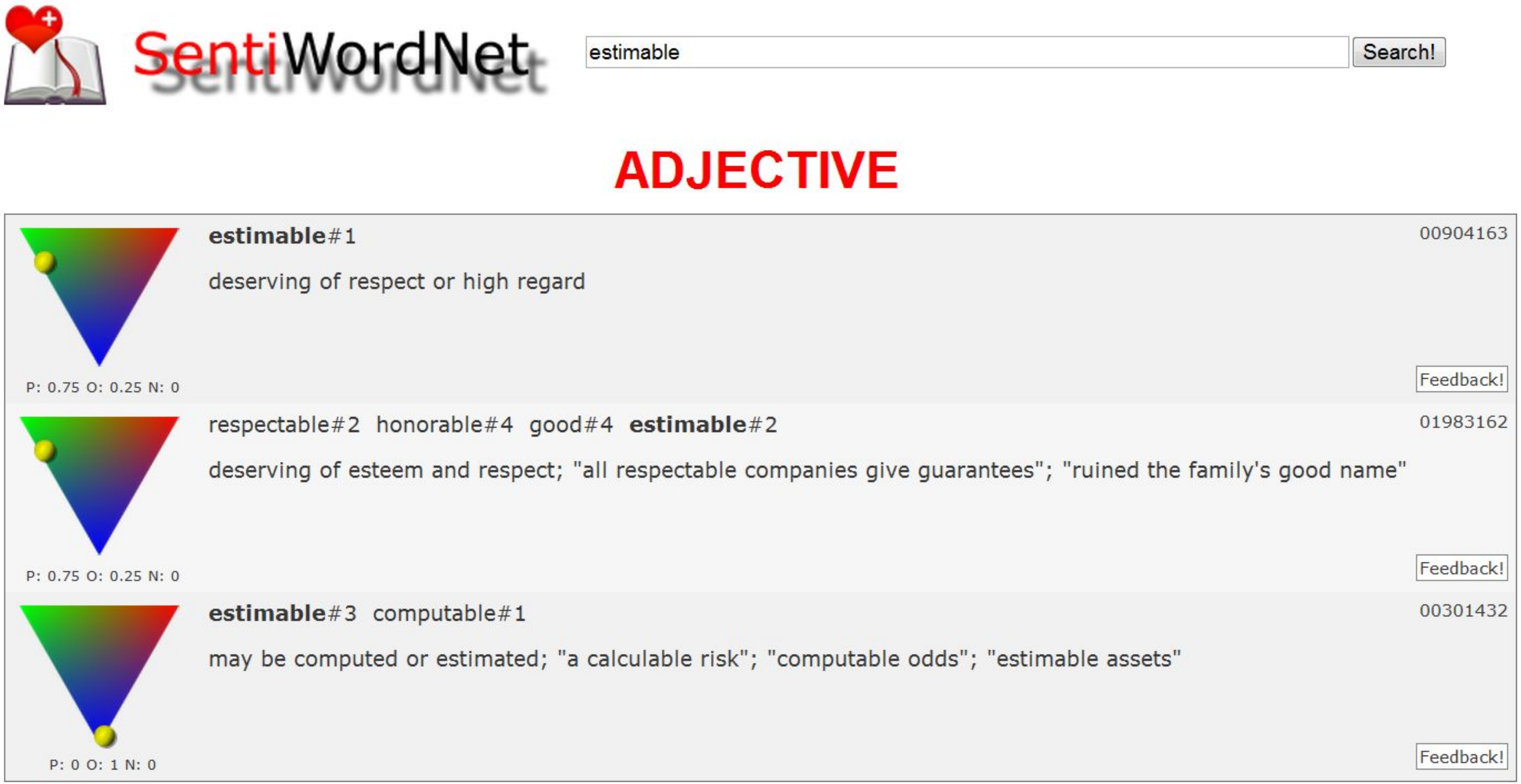


SentiWordNet

Once the algorithm converges, words are ranked by their sentiment.



SentiWordNet



ADJECTIVE

estimable#1

00904163

deserving of respect or high regard

P: 0.75 O: 0.25 N: 0

respectable#2 honorable#4 good#4 **estimable**#2

01983162

deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"

P: 0.75 O: 0.25 N: 0

estimable#3 computable#1

00301432

may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"

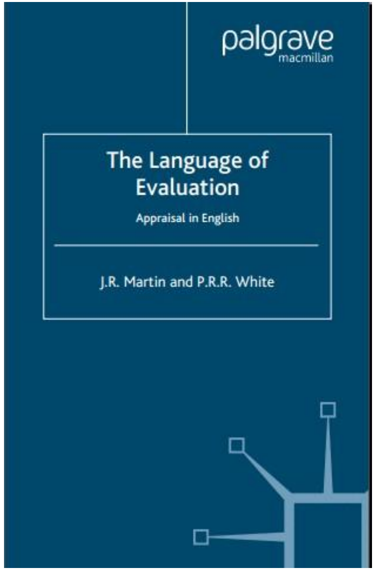
P: 0 O: 1 N: 0

Appraisal theory

The appraisal theory models how evaluation is expressed in text.

The appraisal framework identifies three main components of evaluative language:

- **attitude**: expression of evaluation
  "He is a *good* man"
- **engagement**: who expresses evaluation
  "John *says* he is a good man"
- **graduation**: the strength of the previous two component
  "John *swears* he is a *very* good man"



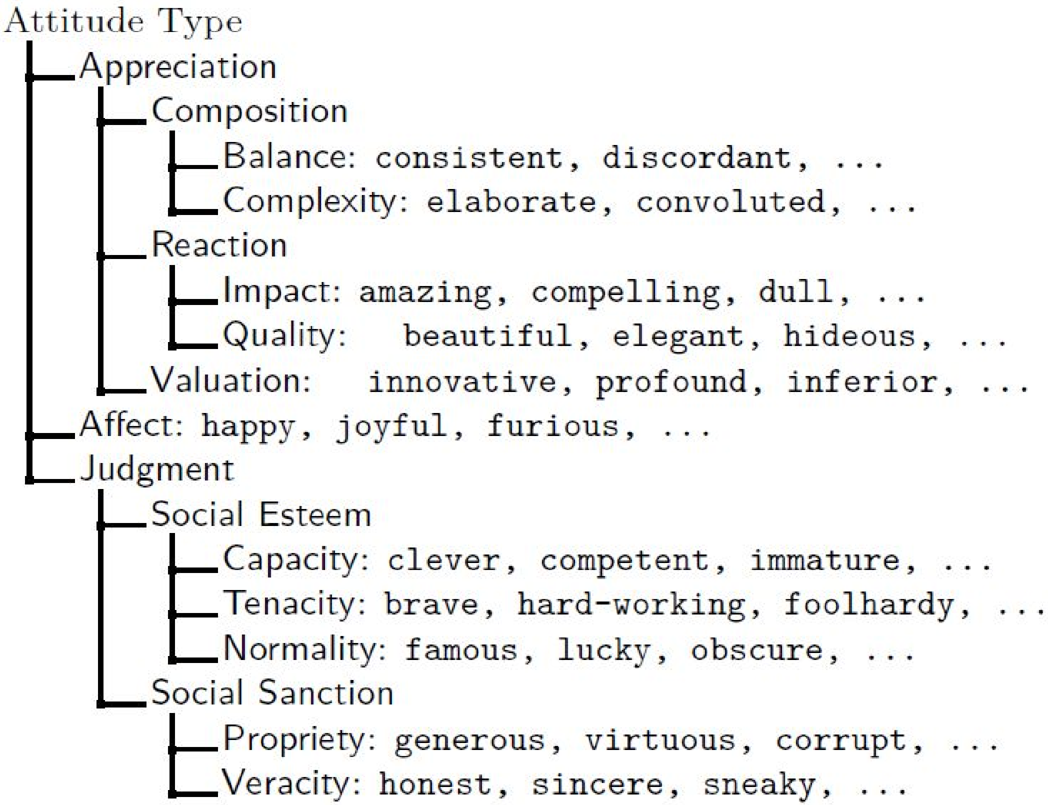
palgrave macmillan

The Language of Evaluation

Appraisal in English

J.R. Martin and P.R.R. White

Appraisal theory



Attitude Type

- Appreciation
  - Composition
    - Balance: consistent, discordant, ...
    - Complexity: elaborate, convoluted, ...
  - Reaction
    - Impact: amazing, compelling, dull, ...
    - Quality: beautiful, elegant, hideous, ...
  - Valuation: innovative, profound, inferior, ...
- Affect: happy, joyful, furious, ...
- Judgment
  - Social Esteem
    - Capacity: clever, competent, immature, ...
    - Tenacity: brave, hard-working, foolhardy, ...
    - Normality: famous, lucky, obscure, ...
  - Social Sanction
    - Propriety: generous, virtuous, corrupt, ...
    - Veracity: honest, sincere, sneaky, ...

Appraisal lexicon

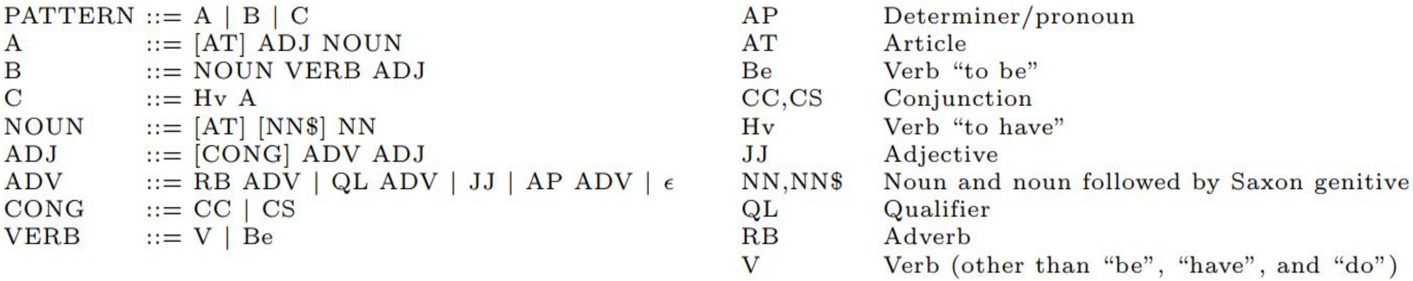
Bloom, Garg and Argamon created an appraisal-focused lexicon.

The lexicon models appraisal properties of 2k words, including *valence shifters*.

```
truly:  {POS:RB, force:increase}

kinda:  {force:decrease}

not:{force:flip, orientation:flip}

nervously:  {POS:RB, attitude:affect,   orientation:negative,
             force:median}

cowardly:   {POS:JJ, attitude:tenacity, orientation:negative,
             force:high, focus:median}
```

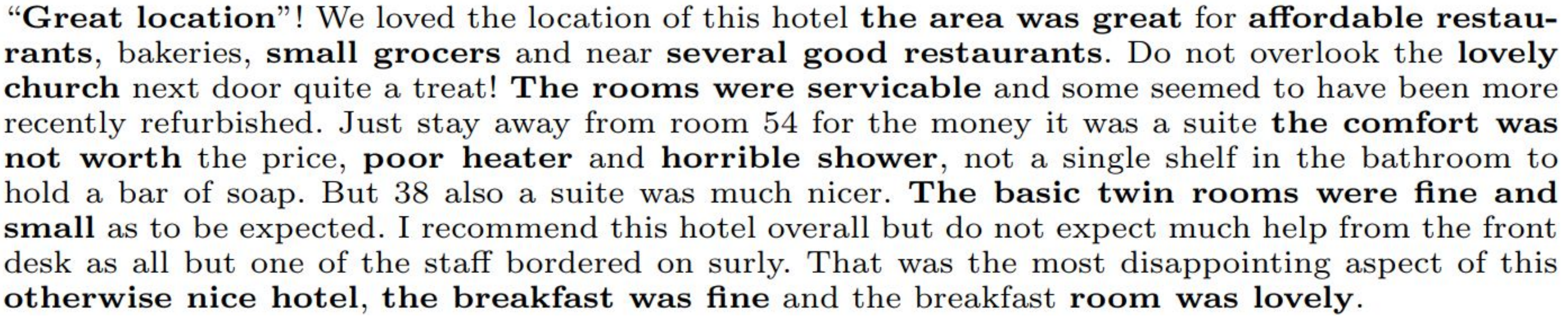
Pattern based features

POS tagging and sentiment lexicons can be combined to extract complex features from text.



```
PATTERN ::= A | B | C
A       ::= [AT] ADJ NOUN
B       ::= NOUN VERB ADJ
C       ::= Hv A
NOUN    ::= [AT] [NN$] NN
ADJ     ::= [CONG] ADV ADJ
ADV     ::= RB ADV | QL ADV | JJ | AP ADV | ϵ
CONG    ::= CC | CS
VERB    ::= V | Be
```

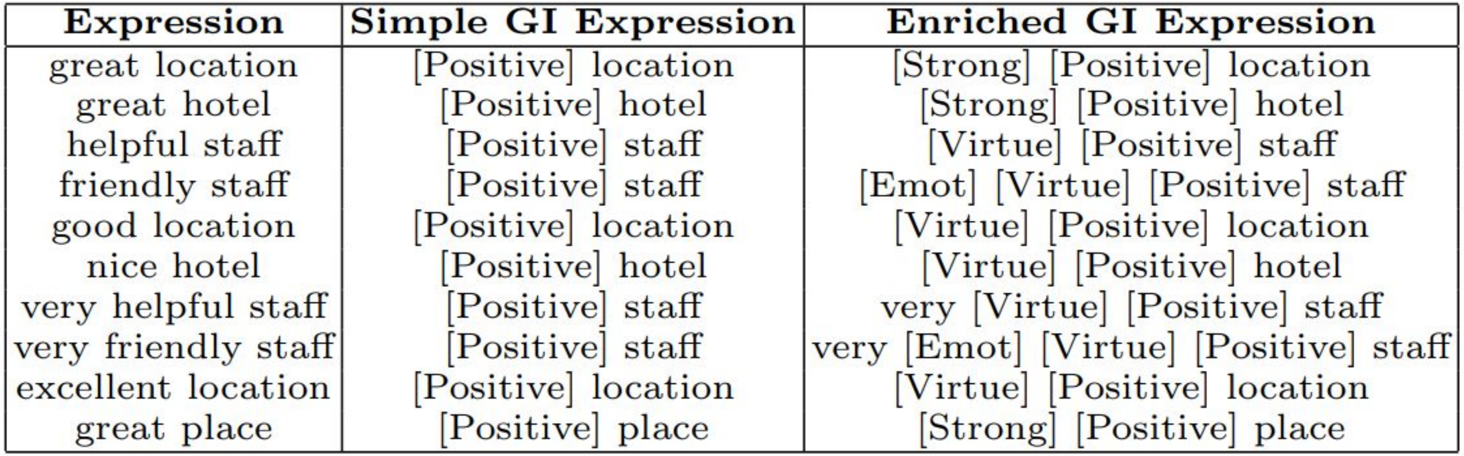
AP	Determiner/pronoun
AT	Article
Be	Verb “to be”
CC,CS	Conjunction
Hv	Verb “to have”
JJ	Adjective
NN,NN$	Noun and noun followed by Saxon genitive
QL	Qualifier
RB	Adverb
V	Verb (other than “be”, “have”, and “do”)



“**Great location**”! We loved the location of this hotel **the area was great** for **affordable restaurants**, bakeries, **small grocers** and near **several good restaurants**. Do not overlook the **lovely church** next door quite a treat! **The rooms were servicable** and some seemed to have been more recently refurbished. Just stay away from room 54 for the money it was a suite **the comfort was not worth** the price, **poor heater** and **horrible shower**, not a single shelf in the bathroom to hold a bar of soap. But 38 also a suite was much nicer. **The basic twin rooms were fine and small** as to be expected. I recommend this hotel overall but do not expect much help from the front desk as all but one of the staff bordered on surly. That was the most disappointing aspect of this **otherwise nice hotel**, **the breakfast was fine** and the breakfast **room was lovely**.

Pattern based features

POS tagging and sentiment lexicons can be combined to extract complex features from text.



Expression	Simple GI Expression	Enriched GI Expression
great location	[Positive] location	[Strong] [Positive] location
great hotel	[Positive] hotel	[Strong] [Positive] hotel
helpful staff	[Positive] staff	[Virtue] [Positive] staff
friendly staff	[Positive] staff	[Emot] [Virtue] [Positive] staff
good location	[Positive] location	[Virtue] [Positive] location
nice hotel	[Positive] hotel	[Virtue] [Positive] hotel
very helpful staff	[Positive] staff	very [Virtue] [Positive] staff
very friendly staff	[Positive] staff	very [Emot] [Virtue] [Positive] staff
excellent location	[Positive] location	[Virtue] [Positive] location
great place	[Positive] place	[Strong] [Positive] place

Baccianella S., Esuli A., Sebastiani F. (2009) Multi-facet Rating of Product Reviews. ECIR 2009. Lecture Notes in Computer Science, vol 5478.

Domain-specific lexicons

Global sentiment lexicons can identify generic signals of positivity or negativity in text.

Every topic and domain has a component of specific elements of language used to express evaluation and sentiment.

"an insane performance" - "an insane behavior"

"a warm welcome" - "a warm beer"

Domain-specific lexicons are usually built starting from a global lexicon and unannotated text from the domain.

"Harsh but **un**fair"

Conjunctions are usually selected according to polarity of the joined words:

nice and sturdy
nice but weak
**ugly but weak*

Build a graph with links determined by the most frequent type of occurrences of conjoined words (and = "same", but = "opposite") in a text collection.

Partition the graph in two parts maximizing "same" links and minimizing "opposite" links inside the partitions.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. EACL 1997.

Pointwise Mutual Information

Domain lexicons can be learned by measuring *statistical correlation* with a selection of *seed terms*, e.g., using PMI:

$$\mathrm{PMI}(w_1, w_2) = \log\left(\frac{p(w_1, w_2)}{p(w_1)p(w_2)}\right)$$

$$\mathrm{SO}(w) = \sum_{\mathrm{seed}_{\mathrm{pos}} \in P, \mathrm{seed}_{\mathrm{neg}} \in N} \log\left(\frac{p(w, \mathrm{seed}_{\mathrm{pos}})p(\mathrm{seed}_{\mathrm{neg}})}{p(w, \mathrm{seed}_{\mathrm{neg}})p(\mathrm{seed}_{\mathrm{pos}})}\right)$$

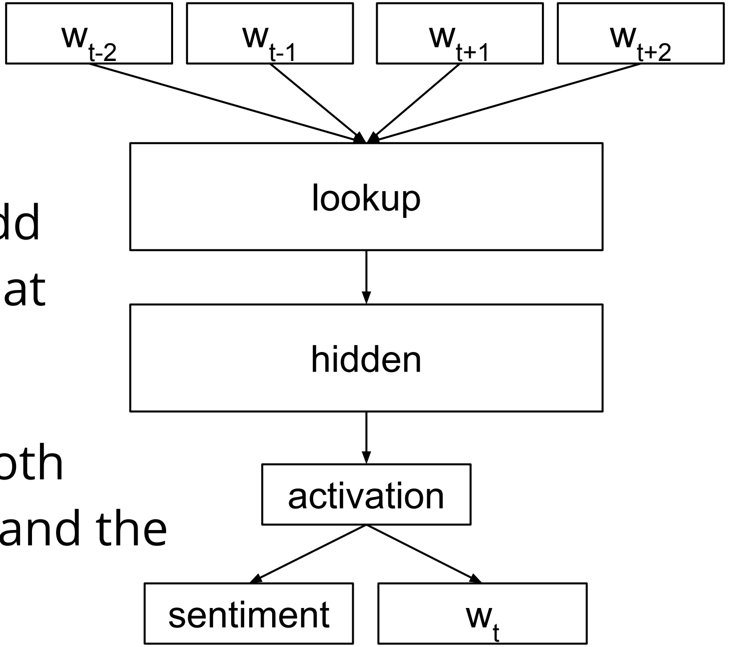
$$\mathrm{P} = \{\textsf{good, nice, excellent, positive, fortunate, correct, superior}\}$$
$$\mathrm{N} = \{\textsf{bad, nasty, poor, negative, unfortunate, wrong, inferior}\}$$

Turney, P., Littman, M. L. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus

Sentiment Embeddings

When labeled data is available it is possible to add sentiment information to the training process that learns the word embeddings.

E.g., extending a CBOW-like neural network to both predict, given a context, the correct target word and the correct sentiment label.



Tang et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. ACL 2014

https://github.com/attardi/deepnl/wiki/Sentiment-Specific-Word-Embeddings

SentProp

SentProp is an algorithm for the annotation of domain-specific sentiment lexicons from an unlabeled text collection.

It uses an algorithm that is similar to the one used for SentiWordNet:

- It starts by computing the word embeddings on the text collection.
- A graph is built based on the similarity between embeddings.
- A set of seed terms defines sources of positivity and negativity.
- A random walk process is run to spread positivity and negativity across the graph.