Sentiment Analysis

Andrea Esuli

What is Sentiment Analysis?

What is Sentiment Analysis?

“Sentiment analysis and opinion mining is the field of study that analyzes people’s opinions, sentiments, evaluations, attitudes, and emotions from written language.”

Bing Liu, “*Sentiment Analysis and Opinion Mining*” Morgan & Claypool Publishers, 2012.

SA works on the **subjective/evaluative/emotive** components of textual information, which have often been ignored in the **objective/factual/topical** analysis usually performed in traditional TA.

Topic vs Sentiment

Topic and sentiment are two main orthogonal dimensions:

- Topic/Fact/Objective information
- Sentiment/Opinion/Subjective information (affective states, emotions. . .)

Topical analysis:

- Discriminating political news from sport news.
- Extracting mention of names of persons in text.

Sentiment analysis:

- Discriminating between favorable and negative attitude toward a subject.
- Identifying the expressions of an emotion and the target of that emotion.

Topic vs Sentiment

Objective information:

*The **4.7-inch** <u>display</u> on the iPhone 6 is arguably its best feature.*

*…concerns have been raised about the relatively low <u>resolution</u> (**1334 x 750 pixels**)*

Source

Topic vs Sentiment

Subjective information:

*The 4.7-inch display on the iPhone 6 is arguably its **best feature**.*

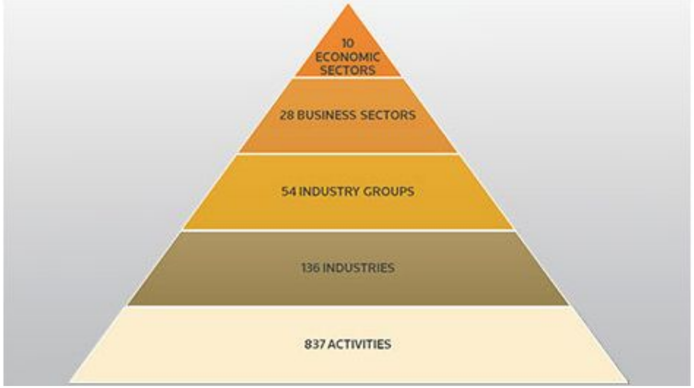
*…concerns have been raised about the **relatively low** resolution (1334 x 750 pixels)*

Source

Topic vs Sentiment

Classification of documents:

- with respect to the Thomson Reuters taxonomy*.



- with respect to the content being a positive, neutral, or a negative evaluation†.

```
{"data": [{"text": "I love Titanic.", "id":1234, "polarity": 4},
          {"text": "I hate Titanic.", "id":4567, "polarity": 0}]}
```

* Source † Source

Topic vs Sentiment

Extraction of information:

- regarding objective properties

  The NBA player Michael Jordan is from the United States of America*

  Organization Person Location

- regarding the expression of opinions.

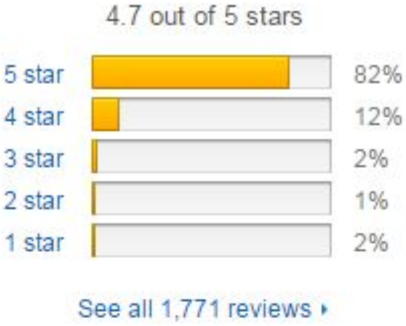
  soldiers with 20 years or more service are generally satisfied with termination packages being offered†

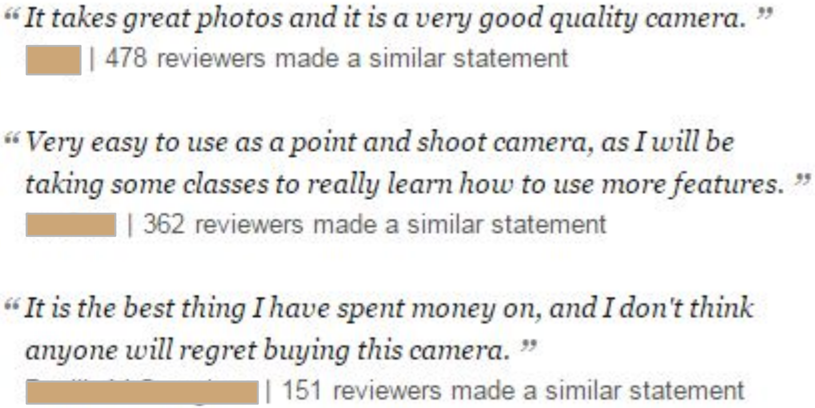
  Agent Attitude Target

* Source † Source

Sentiment and Big Data

Subjective information is varied by definition.
The more sources are compared, the more the vision of the feelings on the matter is complete.

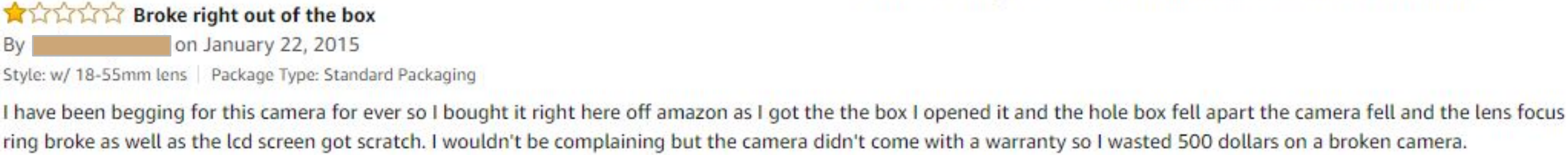




“*It takes great photos and it is a very good quality camera.*”
[redacted] | 478 reviewers made a similar statement

“*Very easy to use as a point and shoot camera, as I will be taking some classes to really learn how to use more features.*”
[redacted] | 362 reviewers made a similar statement

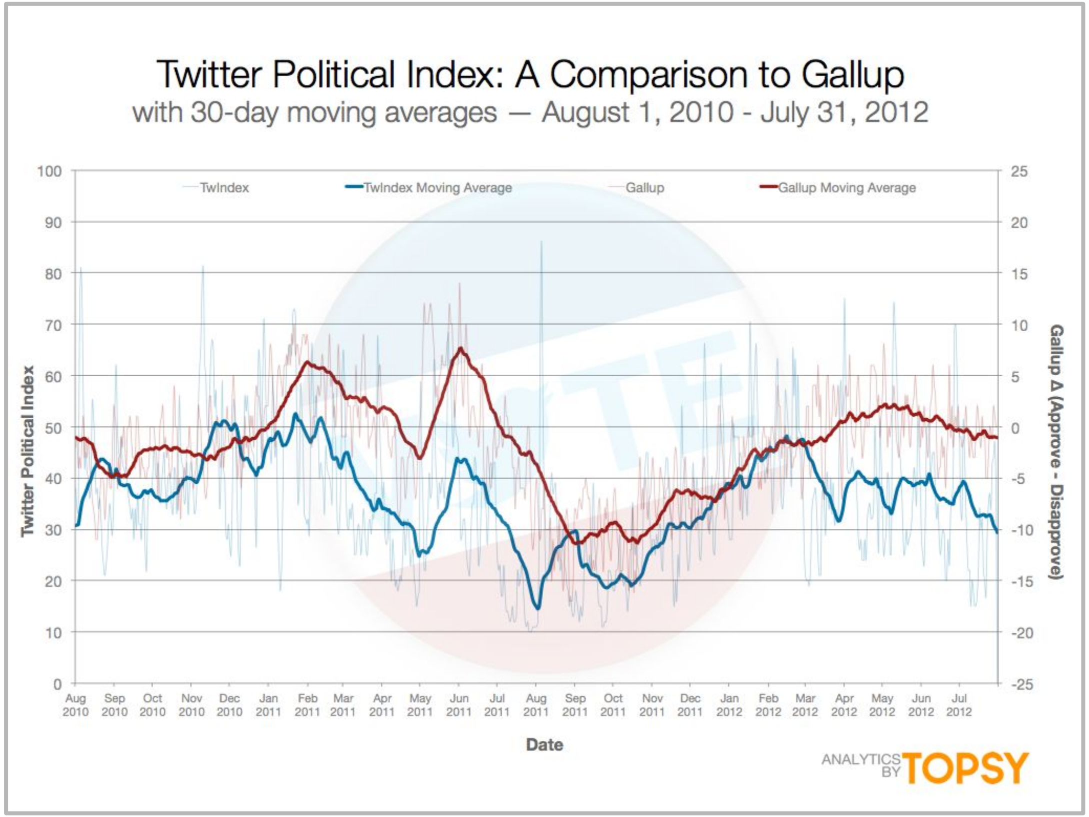
“*It is the best thing I have spent money on, and I don't think anyone will regret buying this camera.*”
[redacted] | 151 reviewers made a similar statement



★☆☆☆☆ **Broke right out of the box**
By [redacted] on January 22, 2015
Style: w/ 18-55mm lens | Package Type: Standard Packaging

I have been begging for this camera for ever so I bought it right here off amazon as I got the the box I opened it and the hole box fell apart the camera fell and the lens focus ring broke as well as the lcd screen got scratch. I wouldn't be complaining but the camera didn't come with a warranty so I wasted 500 dollars on a broken camera.

Sentiment and Big Data



Twindex

Why Sentiment Analysis?

When we have to take a decision we look for the opinion of the others.

The textual user-generated content that is

- shared on the Web/social networks,
- written in open-ended questions in questionnaires,
- sent to companies as feedback, . . .

contains

- **voluntarily produced,**
- **unconstrained,**
- **first-hand/personal,**
- **fresh,**

evaluative information about our topic of interest.

Why Sentiment Analysis?

Practical example: customers satisfaction questionnaires.

- Are you happy with us? yes/no
- How much are you happy on a scale from 0 to 10?
- Your vote is determined by our: □ rates □ service □ other
- Write here any other feedback: ______________________________

The first three answers can be directly **automatically processed** to extract **statistical information**.

The last answer to an **open-ended question** is the only potential source of **unexpected information**.

Sentiment Analysis methods

There is no one-stop solution for Sentiment Analysis.

Sentiment Analysis is not a single problem.
Sentiment Analysis is not a dataset.
Sentiment Analysis is not a lexicon.
Sentiment Analysis is not an algorithm.

Sentiment Analysis is a special scenario for text analysis problems.

A “standard” method produces 70-90% of the result.

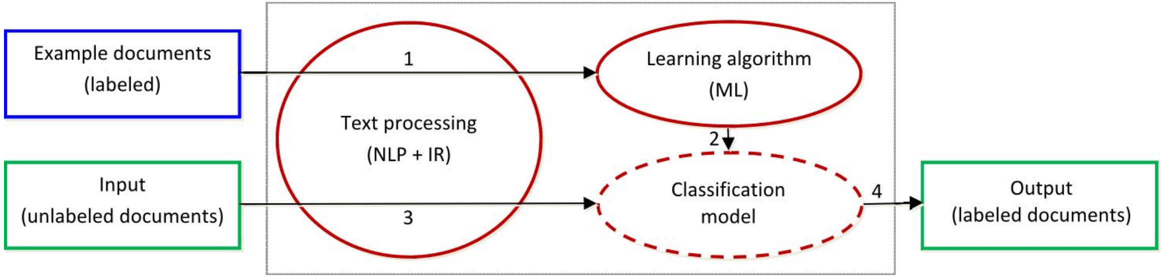
Exploiting the characteristic that are specific of a given Sentiment Analysis problem produces that 10-30% improvement that separates an average solution from a good one.

Sentiment Analysis methods

Multidisciplinary approach:

- Natural Language Processing
- Information Retrieval
- Machine Learning

The template solution to a sentiment analysis problem is the same of a "generic" one, e.g.:


Example documents (labeled)
1
Learning algorithm (ML)
Text processing (NLP + IR)
2
Input (unlabeled documents)
3
Classification model
4
Output (labeled documents)

Most of sentiment-specific methods deal with **capturing how sentiment are expressed in natural language**.

The language of opinions

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.
- Other aspects, e.g., morphology. . .

The main aim of NLP/IR/ML applied to Sentiment Analysis is to **recognize** sentiment expressions and to **model** them into **semantic abstractions**.

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”

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

- The General Inquirer lexicon
- WordNet affect
- SentiWordNet

The language of opinions

Domain/aspect-specific expressions: words that have a sentiment valence only when used in the context of a **specific domain**, or when they are associated with a **specific aspect**.

“The phone is made of cheap plastic”

“The carrier offers cheap rates”

“We have got a warm welcome”

“We have got a warm beer”

A collection of text from the domain can be used to build a **domain lexicon**.

The language of opinions

Negation and **valence shifters**: they do not determine sentiment directly but have **influence** on it.

It is difficult to determine their **scope** and **combined effect**.

"This is a very good car" (increment)

"This car is not very good" (flip, decrement)

"I don't like the design of the new Nokia but it contains some intriguing functions"

"Not only is this phone expensive but it is also heavy and difficult to use"

Workshop on Negation and Speculation in NLP

The language of opinions

Punctuation, emoticons, emoji:

“7AM battery 100% - 9AM 30% :(“

Irony, sarcasm:

“Light as a bulldozer”

“The most useful idea since the DVD rewinder”

Common knowledge:

“Windows Vista: the new Windows ME”

“Windows 7: the new Windows XP”

Affective computing

Modern Sentiment Analysis applications are mainly **data mining oriented** and focused on the evaluations expressed toward the subject matter of the text.

There is also active research on the topic of **affective computing**, more related to **psychology** and **cognitive sciences**.

In affective computing the focus is on the **human computer interaction**, aiming at identifying the **emotions** and **feelings conveyed** by the text **to the reader**.

Affective computing

Recognizing the expression of six basic emotions: anger, disgust, fear, joy, sadness and surprise:

“He looked at his father lying drunk on the floor” (disgust)

“She was leaving and she would never see him again” (sadness)

“She turned and suddenly disappeared from their view” (surprise)

“They celebrated their achievement with an epic party” (joy)

Strapparava and Mihalcea. *Learning to Identify Emotions in Text*. SAC 2008

Computational humor

Generating and recognizing humor: jokes, puns, wordplay.

"Beauty is in the eye of the beholder"
"Beauty is in the eye of the beer holder"

Generation is usually based on templates, recognition is mainly based on stylistic features.

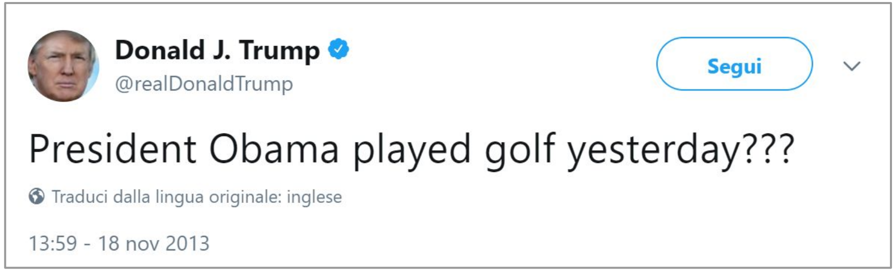
An example of application is building a language playground for people with complex communication needs.

Ritchie et al. *A practical application of computational humour*. ICCC 2007.
Mihalcea and Strapparava. *Learning to Laugh (Automatically): Computational Models for Humor Recognition*. Computational Intelligence, 2006.

Irony and sarcasm

Irony and sarcasm are pervasive on social media.

Both are linguistic phenomena that rely on context and common knowledge.

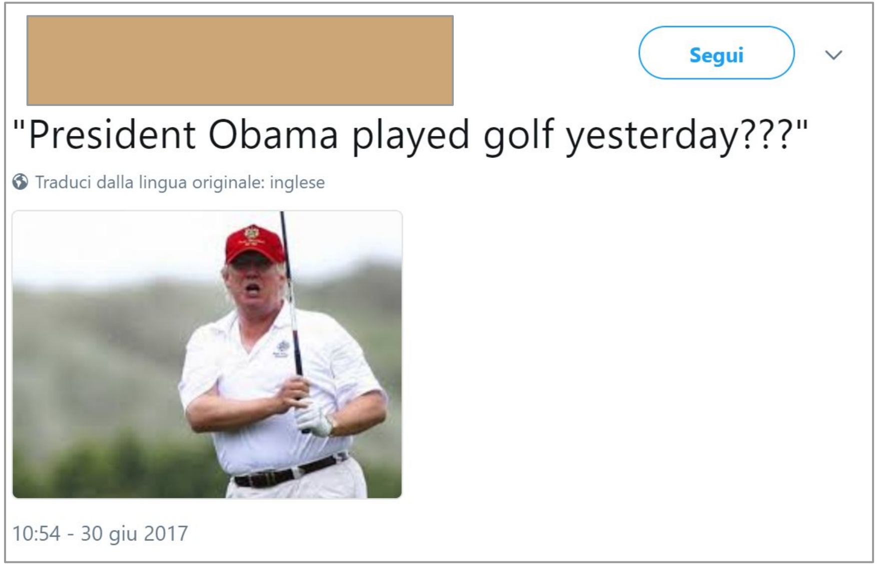


Donald J. Trump
@realDonaldTrump

President Obama played golf yesterday???

Traduci dalla lingua originale: inglese

13:59 - 18 nov 2013



[redacted]

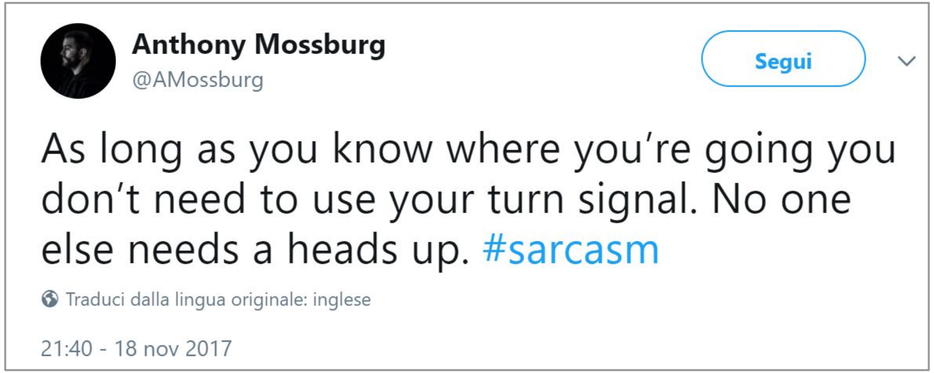
"President Obama played golf yesterday???"

Traduci dalla lingua originale: inglese

10:54 - 30 giu 2017

Irony and sarcasm

Research on computational recognition of irony is at an early stage, mainly focusing on syntactic features.
Data is often collected from tweets with *#ironic* or *#sarcasm* hashtag.



Anthony Mossburg
@AMossburg

As long as you know where you're going you don't need to use your turn signal. No one else needs a heads up. #sarcasm

Traduci dalla lingua originale: inglese

21:40 - 18 nov 2017

[Wallace, "Computational irony: A survey and new perspectives" AIR 2015](#)
[Hernández & Rosso "Irony, Sarcasm, and Sentiment Analysis" Chapter 7 in "Sentiment Analysis in Social Networks" Liu, Messina, Fersini, Pozzi](#)

Lexical resources for Sentiment Analysis

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

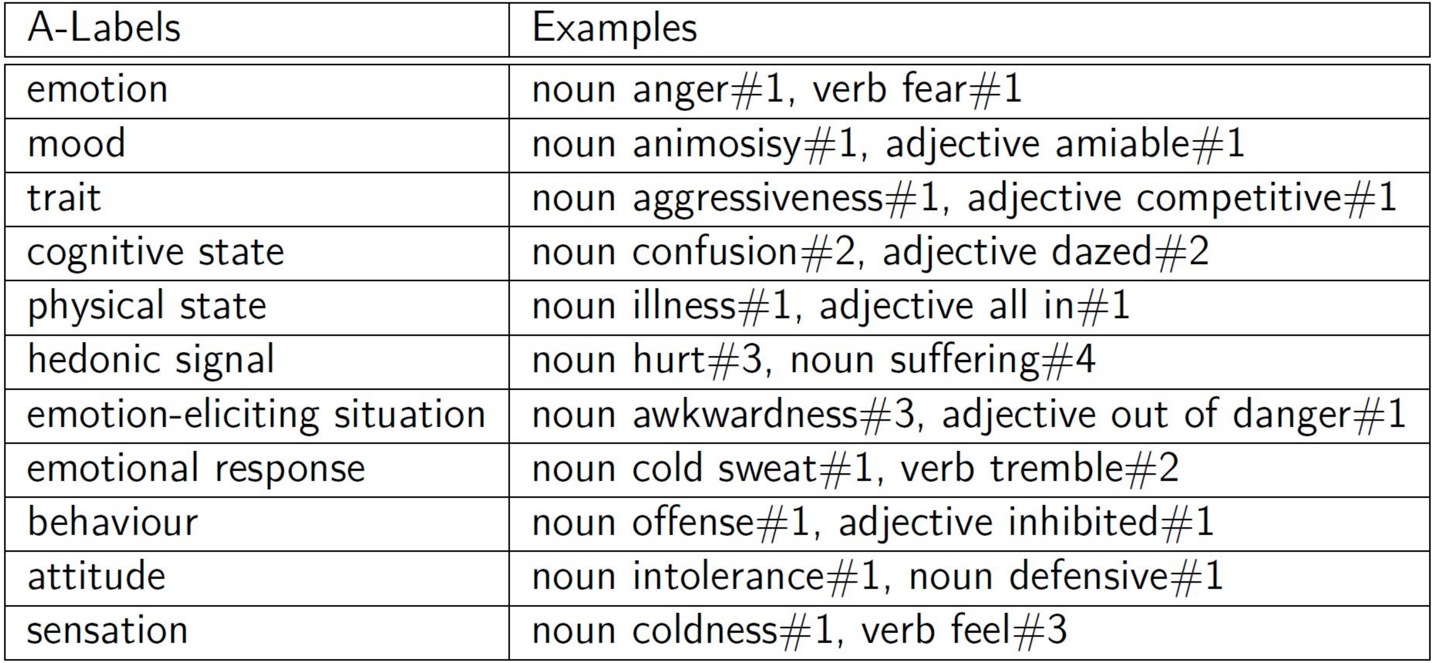
The Multi Perspective Question Answering project produced a number of lexicons for sentiment-related tasks:

- 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

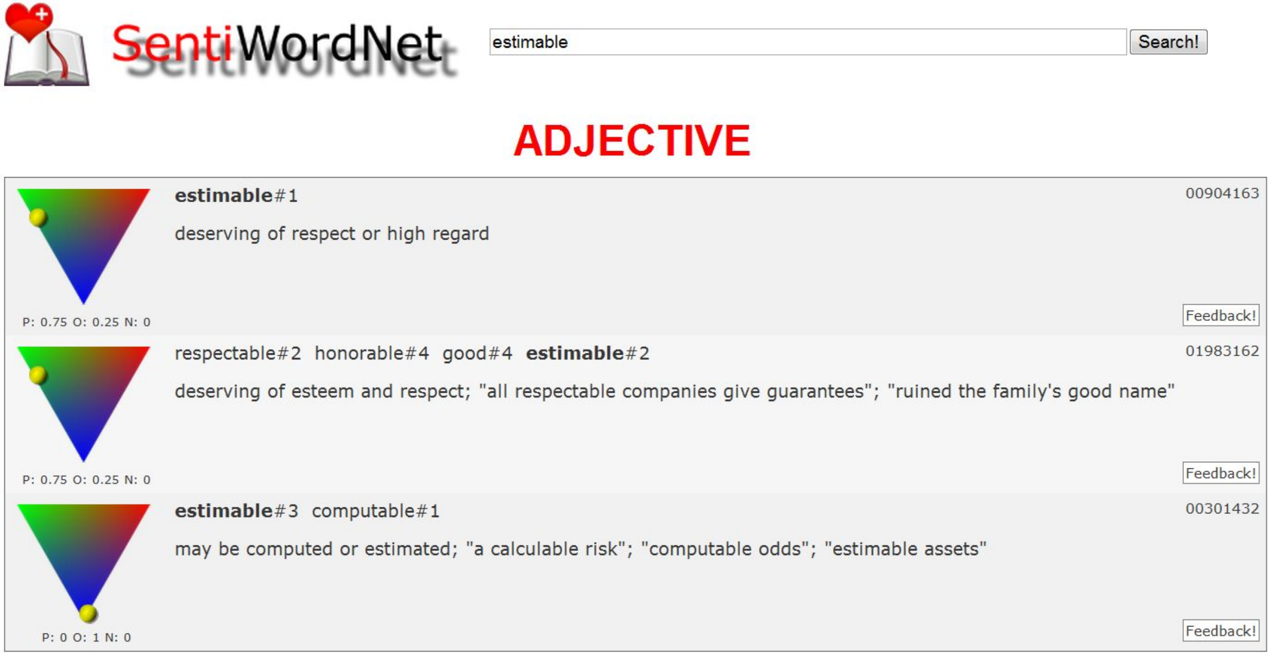
WordNet affect annotates WordNet synsets with emotion-related labels.



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.



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"

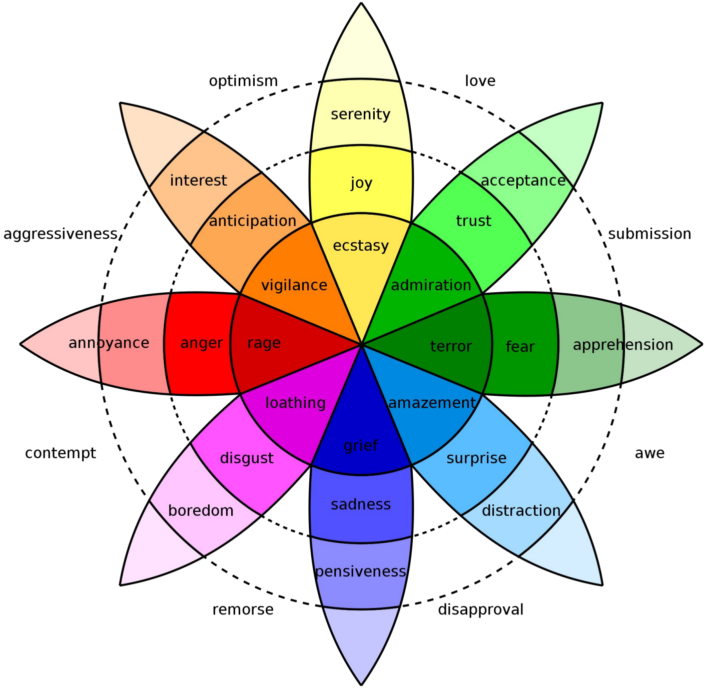
P: 0 O: 1 N: 0

ItEM

ItEM is the Italian EMotional lexicon, that assigns a score related to Plutchik's six basic emotions (***Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise, Anticipation***) to a large set of Italian words.

Starting from 8k+ manually annotated tokens, ItEM uses term similarities based on word embeddings to project the emotion-related properties to new terms.

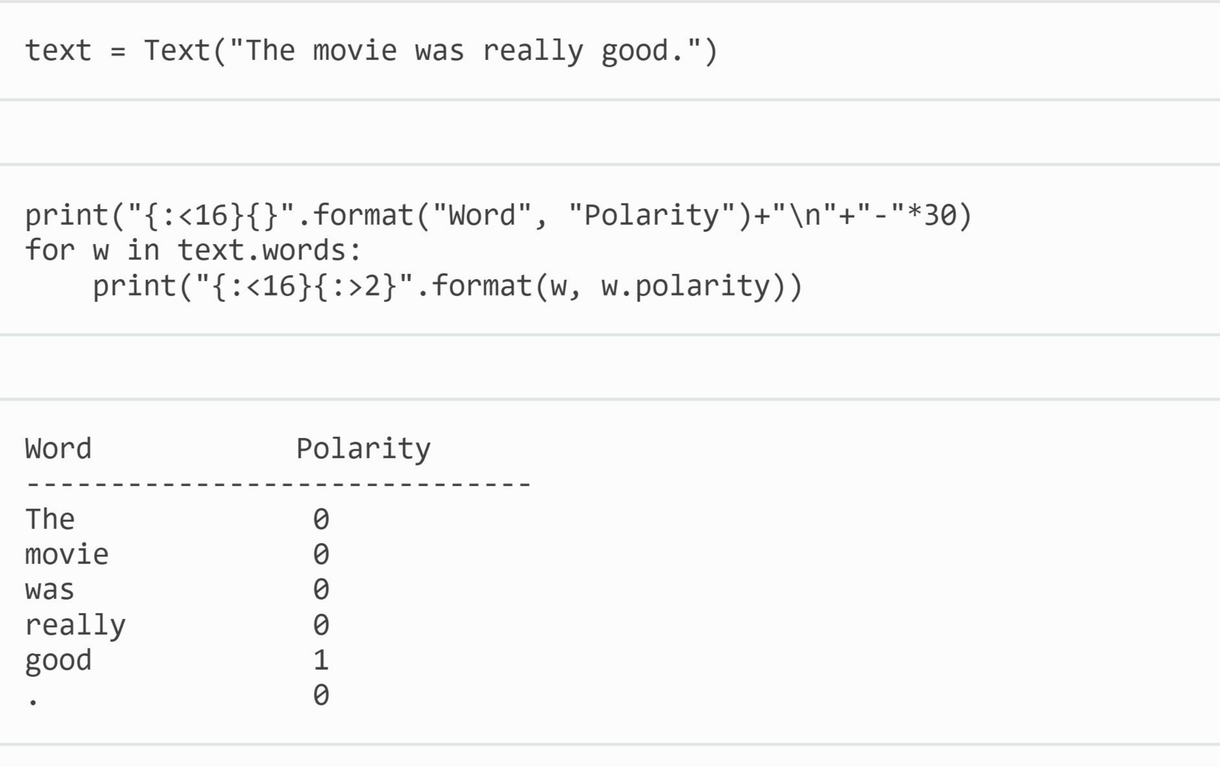
This method allows to create domain-specific lexicons with domain-specific scores.



Polyglot

Polyglot, a python package similar to spaCy, includes sentiment lexicons for +130 languages.

Lexicons are semi-automatically extracted from Wikipedia.



```
text = Text("The movie was really good.")
```

```
print("{:<16}{}".format("Word", "Polarity")+"\n"+"-"*30)
for w in text.words:
    print("{:<16}{:>2}".format(w, w.polarity))
```

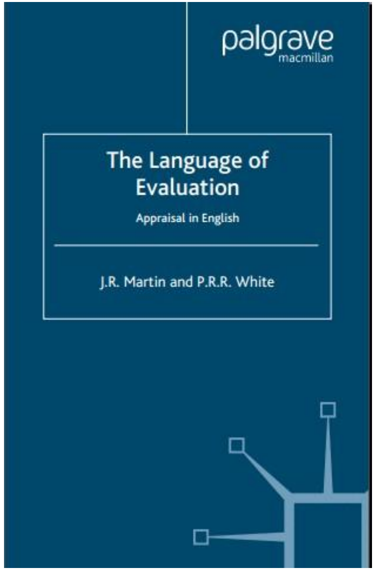
```
Word            Polarity
------------------------------
The              0
movie            0
was              0
really           0
good             1
.                0
```

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"



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}
```

Domain-specific lexicons: "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:

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

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

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

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

Morphology features

Variation in morphology of words is usually reduced/normalized when dealing with topic-oriented analysis:

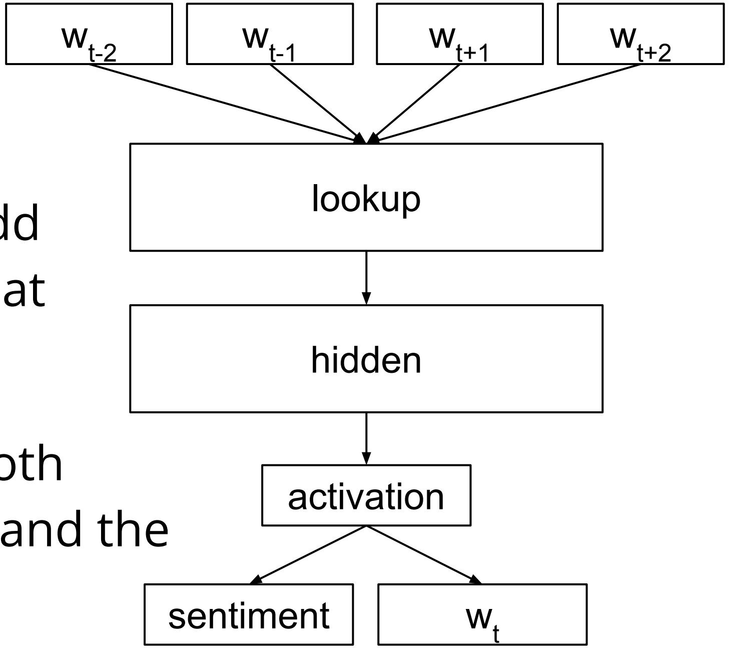
- lowercasing
- reduction to stem/lemma
- removal of misspelled words

The text may exploit word morphology to convey part of the expression of sentiment/opinion.

good, gooood, GOOOOD   great, gr8t, GRRRREAT

Modeling morphology of words with dedicated feature may improve sentiment recognition.

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

Sentiment Classification

Supervised/unsupervised

Supervised learning methods are the most commonly used one, yet also some **unsupervised** methods have been successfully.

Unsupervised methods rely on the shared and recurrent characteristics of the sentiment dimension across topics to perform classification by means of hand-made heuristics and simple language models.

Supervised methods rely on a **training set** of labeled examples that describe the correct classification label to be assigned to a number of documents.
A learning algorithm then exploits the examples to model a general classification function.

Unsupervised methods

Unsupervised Sentiment Classification

Unsupervised methods do not require labeled examples.

Knowledge about the task is usually added by using lexical resources and hard-coded heuristics, e.g.:

- Lexicons + patterns: VADER
- Patterns + Simple language model: SO-PMI

Neural language models have been found that they learn to recognize sentiment with no explicit knowledge about the task.

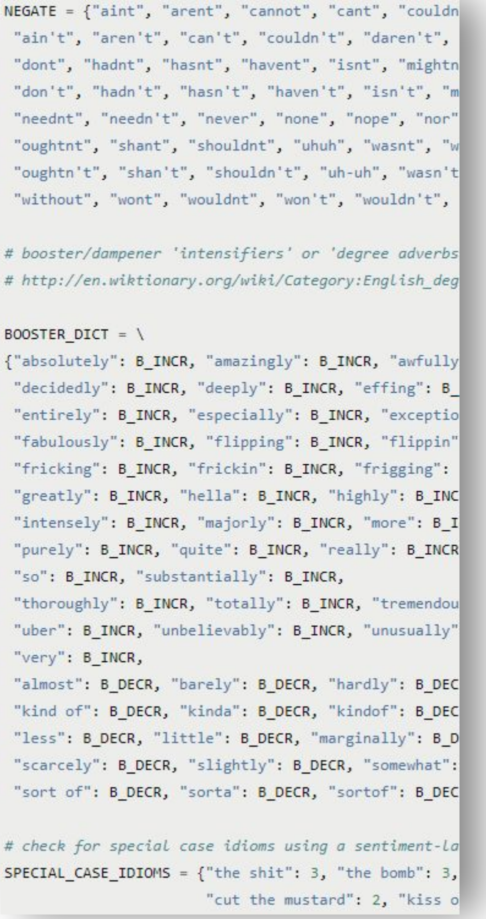
VADER

VADER (Valence Aware Dictionary for sEntiment Reasoning) uses a curated lexicon derived from well known sentiment lexicons that assigns a positivity/negativity score to 7k+ words/emoticons.

It also uses a number of hand-written pattern matching rules (e.g., negation, intensifiers) to modify the contribution of the original word scores to the overall sentiment of text.

Hutto and Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. ICWSM 2014.

VADER is integrated into NLTK



```
NEGATE = {"aint", "arent", "cannot", "cant", "couldn
 "ain't", "aren't", "can't", "couldn't", "daren't",
 "dont", "hadnt", "hasnt", "havent", "isnt", "mightn
 "don't", "hadn't", "hasn't", "haven't", "isn't", "m
 "neednt", "needn't", "never", "none", "nope", "nor"
 "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "w
 "oughtn't", "shan't", "shouldn't", "uh-uh", "wasn't
 "without", "wont", "wouldnt", "won't", "wouldn't",

# booster/dampener 'intensifiers' or 'degree adverbs
# http://en.wiktionary.org/wiki/Category:English_deg

BOOSTER_DICT = \
{"absolutely": B_INCR, "amazingly": B_INCR, "awfully
 "decidedly": B_INCR, "deeply": B_INCR, "effing": B_
 "entirely": B_INCR, "especially": B_INCR, "exceptio
 "fabulously": B_INCR, "flipping": B_INCR, "flippin"
 "fricking": B_INCR, "frickin": B_INCR, "frigging":
 "greatly": B_INCR, "hella": B_INCR, "highly": B_INC
 "intensely": B_INCR, "majorly": B_INCR, "more": B_I
 "purely": B_INCR, "quite": B_INCR, "really": B_INCR
 "so": B_INCR, "substantially": B_INCR,
 "thoroughly": B_INCR, "totally": B_INCR, "tremendou
 "uber": B_INCR, "unbelievably": B_INCR, "unusually"
 "very": B_INCR,
 "almost": B_DECR, "barely": B_DECR, "hardly": B_DEC
 "kind of": B_DECR, "kinda": B_DECR, "kindof": B_DEC
 "less": B_DECR, "little": B_DECR, "marginally": B_D
 "scarcely": B_DECR, "slightly": B_DECR, "somewhat":
 "sort of": B_DECR, "sorta": B_DECR, "sortof": B_DEC

# check for special case idioms using a sentiment-la
SPECIAL_CASE_IDIOMS = {"the shit": 3, "the bomb": 3,
                       "cut the mustard": 2, "kiss o
```

VADER

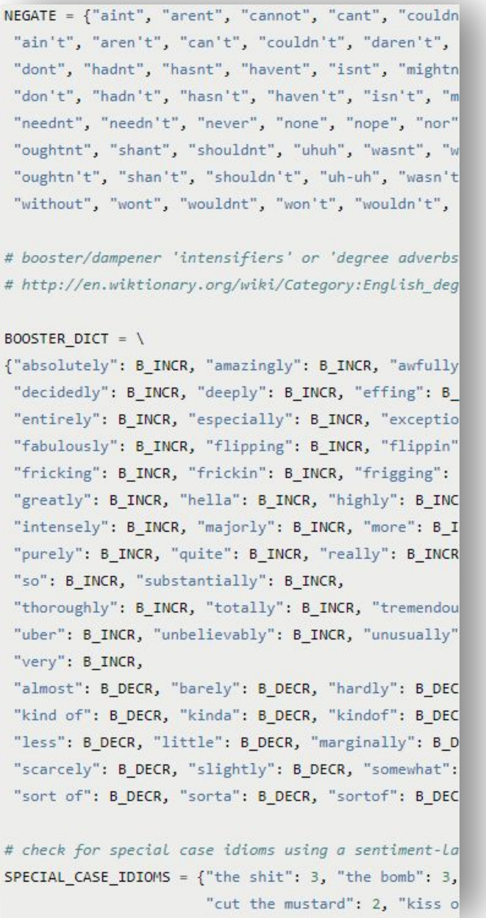
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
vader = SentimentIntensityAnalyzer()

vader.polarity_scores('the best experience I had')
Out: {'neg': 0.0, 'neu': 0.417, 'pos': 0.583, 'compound': 0.6369}

vader.polarity_scores('not the best experience I had')
Out: {'neg': 0.457, 'neu': 0.543, 'pos': 0.0, 'compound': -0.5216}
```

VADER can be used to bootstrap a training set for *supervised learning.*

In this case we can talk of a *weakly-supervised* or *semi-supervised* approach, since training data are not all validated by a human, and can contain errors.



```
NEGATE = {"aint", "arent", "cannot", "cant", "couldn
 "ain't", "aren't", "can't", "couldn't", "daren't",
 "dont", "hadnt", "hasnt", "havent", "isnt", "mightn
 "don't", "hadn't", "hasn't", "haven't", "isn't", "m
 "neednt", "needn't", "never", "none", "nope", "nor"
 "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "w
 "oughtn't", "shan't", "shouldn't", "uh-uh", "wasn't
 "without", "wont", "wouldnt", "won't", "wouldn't",

# booster/dampener 'intensifiers' or 'degree adverbs
# http://en.wiktionary.org/wiki/Category:English_deg

BOOSTER_DICT = \
{"absolutely": B_INCR, "amazingly": B_INCR, "awfully
 "decidedly": B_INCR, "deeply": B_INCR, "effing": B_
 "entirely": B_INCR, "especially": B_INCR, "exceptio
 "fabulously": B_INCR, "flipping": B_INCR, "flippin"
 "fricking": B_INCR, "frickin": B_INCR, "frigging":
 "greatly": B_INCR, "hella": B_INCR, "highly": B_INC
 "intensely": B_INCR, "majorly": B_INCR, "more": B_I
 "purely": B_INCR, "quite": B_INCR, "really": B_INCR
 "so": B_INCR, "substantially": B_INCR,
 "thoroughly": B_INCR, "totally": B_INCR, "tremendou
 "uber": B_INCR, "unbelievably": B_INCR, "unusually"
 "very": B_INCR,
 "almost": B_DECR, "barely": B_DECR, "hardly": B_DEC
 "kind of": B_DECR, "kinda": B_DECR, "kindof": B_DEC
 "less": B_DECR, "little": B_DECR, "marginally": B_D
 "scarcely": B_DECR, "slightly": B_DECR, "somewhat":
 "sort of": B_DECR, "sorta": B_DECR, "sortof": B_DEC

# check for special case idioms using a sentiment-la
SPECIAL_CASE_IDIOMS = {"the shit": 3, "the bomb": 3,
                       "cut the mustard": 2, "kiss o
```

Thumbs Up or Thumbs Down?

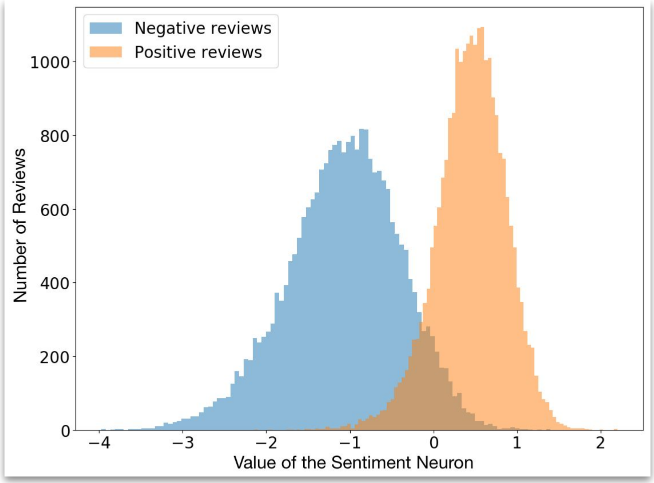
Pointwise Mutual Information has been applied to determine the overall sentiment of text.

- Short phrases extracted from text using POS patterns, e.g.: `JJ+NN, RB+JJ, JJ+JJ, NN+JJ, RB+VB`
- SO-PMI score of each phrase is computed using a search engine and proximity queries, e.g.: `"very solid" NEAR good`
- SO-PMI scores for phrases are averaged to produce the document score.

Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL 2002

Sentiment Classification from a single neuron

A char-level LSTM with 4096 units has been trained on **82 millions** of reviews from Amazon, for text generation.



After training one of the units had a very high correlation with sentiment, resulting in state-of-the-art accuracy when used as a classifier.

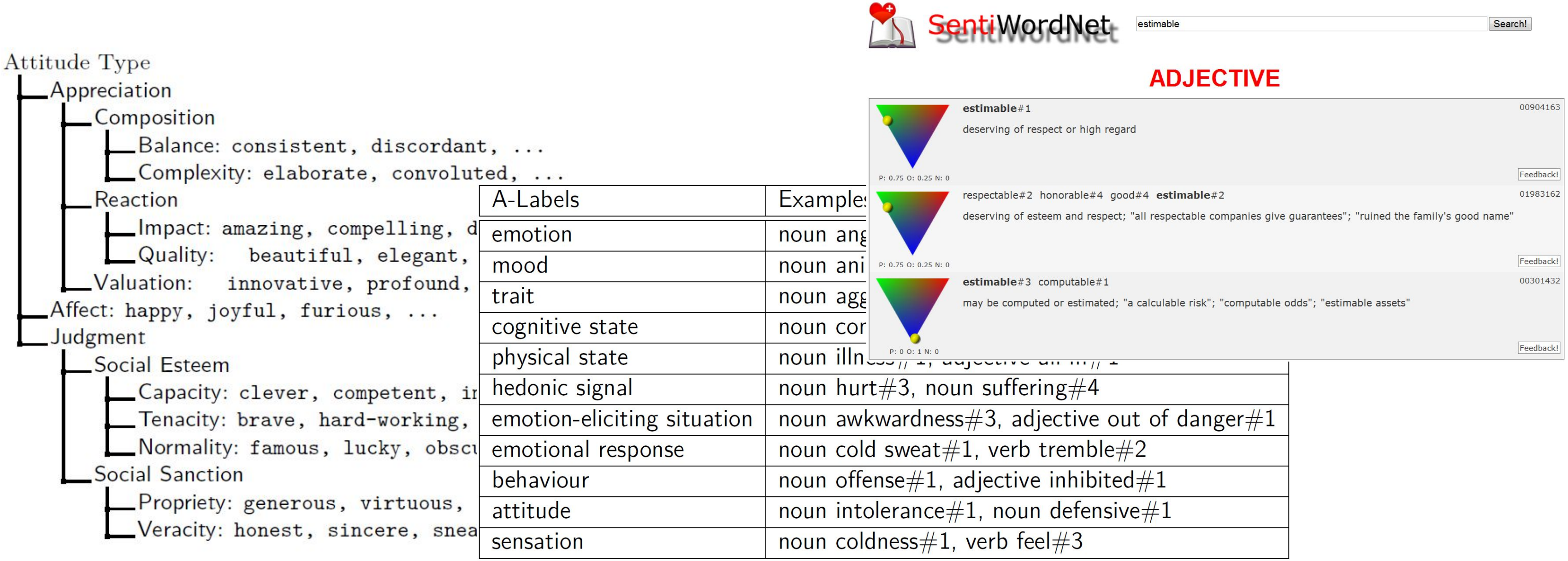
By fixing the sentiment unit to a given value, the generation process has been forced to produce reviews with a given sentiment polarity.

Blog post - Radford et al. Learning to Generate Reviews and Discovering Sentiment. Arxiv 1704.01444

Supervised methods

Supervised methods

Supervised methods use a traditional ML pipeline, typically exploiting the use of lexical resources to improve the number and quality of sentiment-related features extracted from text.



Attitude Type
- Appreciation
  - Composition
    - Balance: consistent, discordant, ...
    - Complexity: elaborate, convoluted, ...
  - Reaction
    - Impact: amazing, compelling, d
    - Quality: beautiful, elegant,
  - Valuation: innovative, profound,
- Affect: happy, joyful, furious, ...
- Judgment
  - Social Esteem
    - Capacity: clever, competent, ir
    - Tenacity: brave, hard-working,
    - Normality: famous, lucky, obscu
  - Social Sanction
    - Propriety: generous, virtuous,
    - Veracity: honest, sincere, snea

A-Labels	Examples
emotion	noun ang
mood	noun ani
trait	noun agg
cognitive state	noun cor
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

estimable

Search!

ADJECTIVE

estimable#1 00904163

deserving of respect or high regard

P: 0.75 O: 0.25 N: 0

Feedback!

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

Feedback!

estimable#3 computable#1 00301432

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

P: 0 O: 1 N: 0

Feedback!

Sentiment features

Sentiment lexicon can be exploited to add sentiment information in text representation.

In this way a general knowledge about language connects words that are observed in the training set with words that occur only in the test set (which would have been considered out-of-vocabulary words).

good → SWN_Pos

gentle → SWN_Pos

bad → SWN_Neg

hostile → SWN_Neg

Distant supervision likes sentiment

Distant supervision fits better with sentiment analysis than with topic-related analysis because in the former it is easier to define negative examples.

A negative sentiment is a concept on its own, opposite to a positive one.

The "negation" of a topic is just the absence of the topic. It is harder to define a heuristic to label negative docs.

- How to automatically mark a negative example for a "soccer" classifier?
    - Just use random sampling when nothing else works.

Distant supervision

Producing training data for supervised learning may have a relevant cost.

Distant supervision exploits "cheap" methods that "weakly" label examples to bootstrap a training set, e.g.:

- labeling tweets with 😄 as positive and those with 😒 as negative.
- using VADER to perform a first labeling (skipping low confidence labels).

The rationale behind distant supervision is that:

- noisy information in training data will cancel out in the learning phase.
- discriminant features that have a decent correlation with the weak labeling emerge among the other.