

SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

Stefano Baccianella, *Andrea Esuli*, Fabrizio Sebastiani

firstname.lastname@isti.cnr.it
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy

LREC 2010,
Malta, May 17–23, 2010

SENTIWORDNET

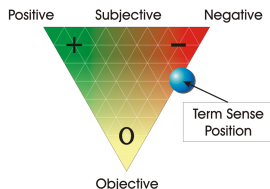
SENTIWORDNET is an **automatically generated** lexical resource that assigns to each synset of WordNet a triple of sentiment-related values: *positivity*, *negativity*, *objectivity*.

SENTIWORDNET has been first presented at LREC 2006, in Genova.

The new SENTIWORDNET 3.0 is aligned to the new WORDNET 3.0.

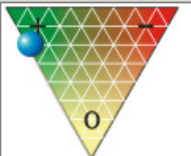
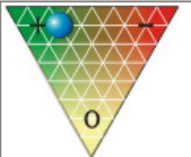
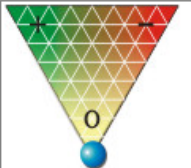
SENTIWORDNET 3.0 is based almost on the same algorithms that generated SENTIWORDNET 1.0 and 2.0.

Enhancement: taking advantage of the **manually sense-disambiguated glosses** available for WORDNET 3.0 (Princeton WordNet Gloss Corpus).



Each synset is assigned with a triple of values that sum up to one.

SENTIWORDNET

 <p>$P = 0.75, N = 0, O = 0.25$</p>	<p><u>estimable(1)</u> <i>deserving of respect or high regard</i></p>
 <p>$P = 0.625, N = 0.25, O = 0.125$</p>	<p><u>honorable(5)</u> <u>good(4)</u> <u>respectable(2)</u> <u>estimable(2)</u> <i>deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</i></p>
 <p>$P = 0, N = 0, O = 1$</p>	<p><u>computable(1)</u> <u>estimable(3)</u> <i>may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</i></p>

Gloss classification: SENTIWORDNET 1.0

Sentiment classification of a WORDNET synset by **classifying its gloss**.

Synset	Gloss
good#a#3	morally admirable
bad#a#1	having undesirable or negative qualities

A **committee of three-way gloss classifiers** (positive/negative/objective) is generated by using a **semi-supervised** learning method.

The training set for a classifier is generated **iteratively**, starting from a **small seed set** of well-known positive, negative, and objective synsets, and adding new synsets by navigating the WORDNET relations.

Each **committee member** uses different parameters (i.e., number of iterations, seed set, learner), making it more or less restrictive in recognizing subjectivity.

The triple of values for a synset is determined as the normalized **count of votes** produced by the committee members for each class.

Gloss classification: SENTIWORDNET 1.0

The classifiers of SENTIWORDNET 1.0 use a traditional **bag of words** model to represent the glosses.

- Ambiguous terms in glosses, e.g., “estimable”, negatively impact on accuracy.

1.0 → 3.0

The classifiers of SENTIWORDNET 3.0 use a **bag of synsets** model to represent the glosses.

The output of this process is SENTIWORDNET 3.0-semi.

Random walk: SENTIWORDNET 2.0

Improving SENTIWORDNET 1.0 by **reassigning values to synsets** based on the output of a **PageRank random walk** algorithm applied to a graph of synsets:

- synsets are the node of the graph;
- a link between a s_i and s_j exists iff s_i appears in the gloss of s_j (definiens \rightarrow definiendum).
 - If a synset is described/pointed mostly by negative synsets it is likely to be negative.
- the PageRank algorithm is used to let positivity flow into the graph, starting from an initial state determined by SENTIWORDNET 1.0 positivity values (that same is separately done for negativity);
- the final PageRank values for positivity and negativity determine how the positivity and negativity values have to be reassigned to synsets.

Random walk: SENTIWORDNET 2.0

EXTENDEDWORDNET is the source of the (automatically) disambiguated glosses for WORDNET 2.0.

Synset	{tidy#v#1, tidy_up#v#1, ... }
WordNet gloss	put (things or places) in order;
eXtendedWordNet gloss	put#v#1 (things#n#1 or places#n#6) in order#n#15

2.0 → 3.0

The manually disambiguated glosses are a more **reliable and complete** resource than EXTENDEDWORDNET.

- The currently available release of EXTENDEDWORDNET does not disambiguate the glosses of adverbs.
 - We put links for all the senses.

The source for the initial values of the random walk algorithm is SENTIWORDNET 3.0-semi, instead of SENTIWORDNET 1.0.

Evaluation

Micro-WN(Op) is a corpus of 1105 human annotated synsets, using the same annotation of model SENTIWORDNET.

Issue: Micro-WN(Op) is aligned to WORDNET 2.0.

We have **automatically** mapped it to WORDNET 3.0 (Micro-WN(Op)-3.0) by using the publicly available **synset mappings** (available only for nouns and verbs) and a **gloss similarity**-based mapping **heuristic**.

The various SENTIWORDNET versions are evaluated by comparing how they rank the synsets of Micro-WN(Op) by positivity, or negativity, with respect to the ranking determined by human annotators.

Evaluation measure: ***p*-normalized Kendall τ distance**

$$\tau_p = \frac{n_d + p \cdot n_u}{Z} \quad (1)$$

Lower values indicate higher agreement.

Evaluation

	Rankings	
	Positivity	Negativity
SENTIWORDNET 1.0	.349	.296
SENTIWORDNET 2.0	.292	.222
SENTIWORDNET 3.0-semi	.339	.286
SENTIWORDNET 3.0	.281	.231

Table 1: τ_p values for the positivity and negativity rankings derived from SENTIWORDNET 1.0, 2.0, 3.0-semi, and 3.0, as measured on Micro-WN(Op) and Micro-WN(Op)-3.0.

SENTIWORDNET 3.0-semi improves over SENTIWORDNET 1.0.

The relative improvement of SENTIWORDNET 3.0 over SENTIWORDNET 1.0 is -19.48% for positivity and -21.96% for negativity.

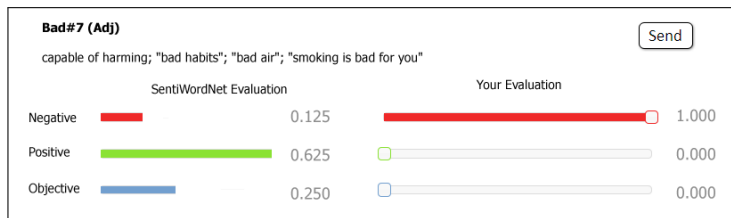
SENTIWORDNET 2.0 obtains a better result on negativity, but SENTIWORDNET 3.0 results are better balanced.

Online user's feedback

SENTIWORDNET is generated by an automated process, it contains errors. It is common for a paper using SENTIWORDNET to report some of such errors.

‘‘for the term bad there is an entry with pos=0, neg=1, obj=0 and another entry with pos = 0.625, neg = 0.125, obj = 0.25 which are completely conflictive’’ [Denecke, 2009]

Collecting user feedback, why not?



User feedback will be released as public domain.

Conclusion

SENTIWORDNET 3.0 and Micro-WN(Op)-3.0 are available at:

<http://swn.isti.cnr.it/>

SENTIWORDNET 3.0 improves over the previous SENTIWORDNET versions:

- by using a bag-of-synsets for gloss representation in the semi-supervised learning step;
- by using manually disambiguated glosses in the random walk step.

The evaluation of SENTIWORDNET 3.0 is based on a gold standard that has been automatically aligned to WORDNET 3.0.

- Adjectives and adverbs have been mapped by using a gloss similarity heuristic.

Collection of user feedback will allow to improve SENTIWORDNET and to develop a dedicated gold standard for WORDNET 3.0.

Thank you. Questions?

