



Resources and Dataset Descriptions

The 2024 Tidelift

impact maintainer

report now on

AND TESTING RESOURCES (items 4-12):

reduced OSS risk [Read](#)

now!

1. vader_icwsm2014_final.pdf

tidelift.com

The original paper for the data set, see citation information (above).

Ads by EthicalAds

2. vader_lexicon.txt

FORMAT: the file is tab delimited with TOKEN, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-HUMAN-SENTIMENT-RATINGS

NOTE: The current algorithm makes immediate use of the first two elements (token and mean valence). The final two elements (SD and raw ratings) are provided for rigor. For example, if you want to follow the same rigorous process that we used for the study, you should find 10 independent humans to evaluate/rate each new token you want to add to the lexicon, make sure the standard deviation doesn't exceed 2.5, and take the average rating for the valence. This will keep the file consistent.

DESCRIPTION: Empirically validated by multiple independent human judges, VADER incorporates a "gold-standard" sentiment lexicon that is especially attuned to microblog-like contexts.

The VADER sentiment lexicon is sensitive both the **polarity** and the **intensity** of sentiments expressed in social media contexts, and is also generally applicable to sentiment analysis in other domains.

Sentiment ratings from 10 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability). Over 9,000 token features were rated on a scale from "[−4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". We kept every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of those ten independent raters. This left us with just over 7,500 lexical features with validated valence scores that indicated both the sentiment polarity (positive/negative), and the sentiment intensity on a scale from −4 to +4. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "latest" is 3.1, whereas "horrible" is −2.5, the frowning emoticon :(is −2.2, and "sucks" and its slang derivative "sux" are both −1.5.

 [latest](#)

Manually creating (much less, validating) a comprehensive sentiment lexicon is a labor intensive and sometimes error prone process, so it is no wonder that many opinion mining researchers and practitioners rely so heavily on existing lexicons as primary resources. We are pleased to offer ours as a new resource. We began by constructing a list inspired by examining existing well-established sentiment word-banks (LIWC, ANEW, and GI). To this, we next incorporate numerous lexical features common to sentiment expression in microblogs, including:

- a full list of Western-style emoticons, for example, :-) denotes a smiley face and generally indicates positive sentiment
- sentiment-related acronyms and initialisms (e.g., LOL and WTF are both examples of sentiment-laden initialisms)
- commonly used slang with sentiment value (e.g., nah, meh and giggly).

We empirically confirmed the general applicability of each feature candidate to sentiment expressions using a wisdom-of-the-crowd (WotC) approach (Surowiecki, 2004) to acquire a valid point estimate for the sentiment valence (polarity & intensity) of each context-free candidate feature.

3. vaderSentiment.py


The Python code for the rule-based sentiment analysis engine. Implements the grammatical and syntactical rules described in the paper, incorporating empirically derived quantifications for the impact of each rule on the perceived intensity of sentiment in sentence-level text. Importantly, these heuristics go beyond what would normally be captured in a typical bag-of-words model. They incorporate **word-order sensitive relationships** between terms. For example, degree modifiers (also called intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity. Consider these examples:

- a. "The service here is extremely good"
- b. "The service here is good"
- c. "The service here is marginally good"

From Table 3 in the paper, we see that for 95% of the data, using a degree modifier increases the positive sentiment intensity of example (a) by 0.227 to 0.36, with a mean difference of 0.293 on a rating scale from 1 to 4. Likewise, example (c) reduces the perceived sentiment intensity by 0.293, on average.

4. tweets_GroundTruth.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, and TWEET-TEXT

DESCRIPTION: includes "tweet-like" text as inspired by 4,000 tweets pulled from  [latest](#) Twitter's public timeline, plus 200 completely contrived tweet-like texts intended to specifically test syntactical and grammatical conventions of conveying differences in

sentiment intensity. The “tweet-like” texts incorporate a fictitious username (@anonymous) in places where a username might typically appear, along with a fake URL ([http://url_removed](#)) in places where a URL might typically appear, as inspired by the original tweets. The ID and MEAN-SENTIMENT-RATING correspond to the raw sentiment rating data provided in ‘tweets_anonDataRatings.txt’ (described below).

5. tweets_anonDataRatings.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-SENTIMENT-RATINGS

DESCRIPTION: Sentiment ratings from a minimum of 20 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability).

6. nytEditorialSnippets_GroundTruth.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, and TEXT-SNIPPET

DESCRIPTION: includes 5,190 sentence-level snippets from 500 New York Times opinion news editorials/articles; we used the NLTK tokenizer to segment the articles into sentence phrases, and added sentiment intensity ratings. The ID and MEAN-SENTIMENT-RATING correspond to the raw sentiment rating data provided in ‘nytEditorialSnippets_anonDataRatings.txt’ (described below).

7. nytEditorialSnippets_anonDataRatings.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-SENTIMENT-RATINGS

DESCRIPTION: Sentiment ratings from a minimum of 20 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability).

8. movieReviewSnippets_GroundTruth.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, and TEXT-SNIPPET

DESCRIPTION: includes 10,605 sentence-level snippets from rotten.tomatoes.com. The snippets were derived from an original set of 2000 movie reviews (1000 positive and 1000 negative) in Pang & Lee (2004); we used the NLTK tokenizer to segment the reviews into sentence phrases, and added sentiment intensity ratings. The ID and MEAN-SENTIMENT-RATING correspond to the raw sentiment rating data provided in ‘movieReviewSnippets_anonDataRatings.txt’ (described below).

9. movieReviewSnippets_anonDataRatings.txt

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-SENTIMENT-RATINGS



latest

DESCRIPTION: Sentiment ratings from a minimum of 20 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability).

10. **amazonReviewSnippets_GroundTruth.txt**

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, and TEXT-SNIPPET

DESCRIPTION: includes 3,708 sentence-level snippets from 309 customer reviews on 5 different products. The reviews were originally used in Hu & Liu (2004); we added sentiment intensity ratings. The ID and MEAN-SENTIMENT-RATING correspond to the raw sentiment rating data provided in 'amazonReviewSnippets_anonDataRatings.txt' (described below).

11. **amazonReviewSnippets_anonDataRatings.txt**

FORMAT: the file is tab delimited with ID, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-SENTIMENT-RATINGS

DESCRIPTION: Sentiment ratings from a minimum of 20 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability).

12. **Comp.Social website with more papers/research:**

[Comp.Social](<http://comp.social.gatech.edu/papers/>)