Rules for sentiment classification

CS 685, Spring 2020

Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

https://people.cs.umass.edu/~brenocon/cs685_s20/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- Diff topic at test/runtime than at training time

  Domain shift

- Collect labels for training — hard!

- When can you NOT use machine learning to do NLP?

classif.

- Problems with classif. in general

VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

C.J. Hutto **Eric Gilbert**
Georgia Institute of Technology, Atlanta, GA 30032
cjhutto@gatech.edu gilbert@cc.gatech.edu

Open-source

- [Hutto and Gilbert, *ICWSM 2014]*
- Sentiment classification WITHOUT requiring labels (rule-based classifier)
  - Sentiment lexicon from crowdworkers
  - Context-aware selection/counting rules
  - Outperforms ML classifiers cross-domain!
- Big contrast to *Pang et al. (2002)* that we talked about before!

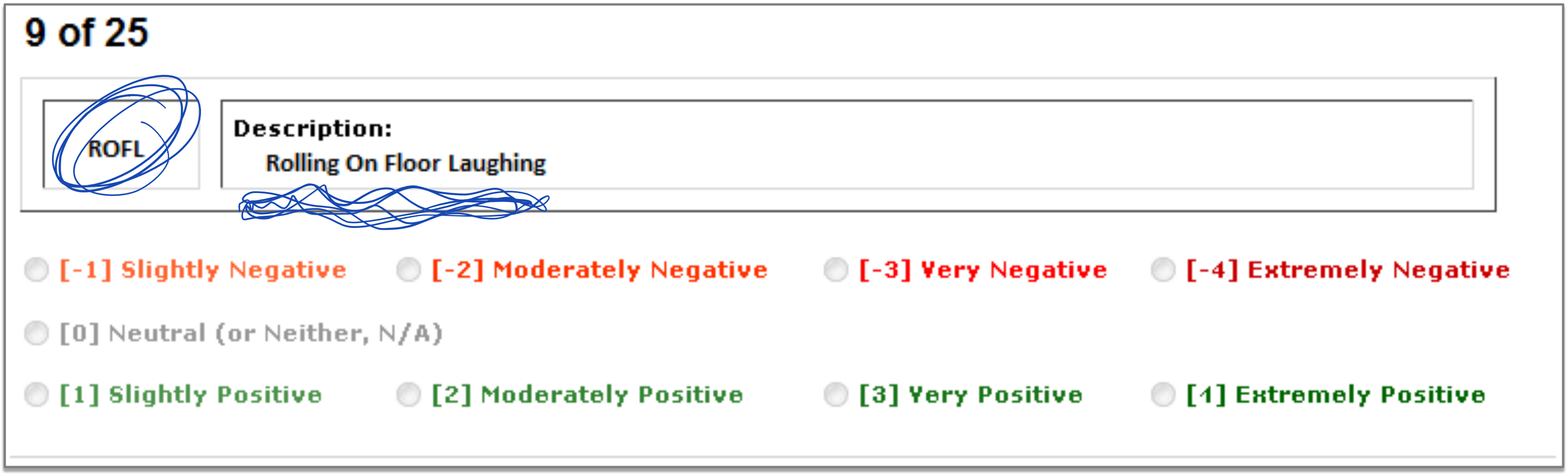

Figure 2: Example of the interface implemented for acquiring valid point estimates of sentiment valence (intensity) for each context-free candidate feature comprising the VADER sentiment lexicon. A similar UI was used for all rating activities described in sections 3.1-3.4.

Crowdsourcing: quality control is hard!

3.1.1 Screening, Training, Selecting, and Data Quality Checking Crowd-Sourced Evaluations and Validations

Previous linguistic rating experiments using a WotC approach on AMT have shown to be reliable – sometimes even outperforming expert raters (Snow, O'Connor, Jurafsky, & Ng, 2008). On the other hand, prior work has also advised on methods to reduce the amount of noise from AMT workers who may produce poor quality work (Downs, Holbrook, Sheng, & Cranor, 2010; Kittur, Chi, & Suh, 2008). We therefore implemented four quality control processes to help ensure we received meaningful data from our AMT raters.

First, every rater was prescreened for English language reading comprehension – each rater had to individually score an 80% or higher on a standardized college-level reading comprehension test.

Second, every prescreened rater then had to complete an online sentiment rating training and orientation session, and score 90% or higher for matching the known (pre-validated) mean sentiment rating of lexical items which included individual words, emoticons, acronyms, sentences, tweets, and text snippets (e.g., sentence segments, or phrases). The user interface employed during the sentiment training (Figure 2) always matched the specific sentiment rating tasks discussed in this paper. The training helped to ensure consistency in the rating rubric used by each independent rater.

Third, every batch of 25 features contained five "golden items" with a known (pre-validated) sentiment rating distribution. If a worker was more than one standard deviation away from the mean of this known distribution on three or more of the five golden items, we discarded all 25 ratings in the batch from this worker.

Finally, we implemented a bonus program to incentivize and reward the highest quality work. For example, we asked workers to select the valence score that they thought "*most other people*" would choose for the given lexical feature (early/iterative pilot testing revealed that wording the instructions in this manner garnered a much tighter standard deviation without significantly affecting the mean sentiment rating, allowing us to achieve higher quality (generalized) results while being more economical).

Heuristics for word matching

- Use context and orthography to better understand a word's impact on text sentiment

1. Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation. For example, "*The food here is good!!!*" is more intense than "*The food here is good*."
2. Capitalization, specifically using ALL-CAPS to emphasize a sentiment-relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation. For example, "*The food here is GREAT!*" conveys more intensity than "*The food here is great!*"
3. Degree modifiers (also called *intensifiers*, *booster words*, or *degree adverbs*) impact sentiment intensity by either increasing or decreasing the intensity. For example, "*The service here is extremely good*" is more intense than "*The service here is good*", whereas "*The service here is marginally good*" reduces the intensity.
4. The contrastive conjunction "*but*" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. "*The food here is great, but the service is horrible*" has mixed sentiment, with the latter half dictating the overall rating.
5. By examining the tri-gram preceding a sentiment-laden lexical feature, we catch nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be "*The food here isn't really all that great*".

Evaluating heuristics

- It's ok to make up examples for carefully controlled tests!

3.3 Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics

Using the general heuristics we just identified, we next selected 30 baseline tweets and manufactured six to ten variations of the exact same text, controlling the specific grammatical or syntactical feature that is presented as an independent variable in a small experiment. With all of the

Test Condition	Example Text
Baseline	Yay. Another good phone interview.
Punctuation1	Yay! Another good phone interview!
Punctuation1 + Degree Mod.	Yay! Another extremely good phone interview!
Punctuation2	Yay!! Another good phone interview!!
Capitalization	YAY. Another GOOD phone interview.
Punct1 + Cap.	YAY! Another GOOD phone interview!
Punct2 + Cap.	YAY!! Another GOOD phone interview!!
Punct3 + Cap.	YAY!!! Another GOOD phone interview!!!
Punct3 + Cap. + Degree Mod.	YAY!!! Another EXTREMELY GOOD phone interview!!!

Table 2: Example of baseline text with eight test conditions comprised of grammatical and syntactical variations.

Table 3 shows the *t*-test statistic, *p*-value, mean of differences for rank ordered data points between each distribution, and 95% confidence intervals:

Test Condition	*t*	*p*	Diff.	95% C.I.
Punctuation (. vs !)	19.02	< 2.2e-16	0.291	0.261 - 0.322
Punctuation (! vs !!)	16.53	2.7e-16	0.215	0.188 - 0.241
Punctuation (!! vs !!!)	14.07	1.7e-14	0.208	0.178 - 0.239
All CAPS (w/o vs w)	28.95	< 2.2e-16	0.733	0.682 - 0.784
Deg. Mod. (w/o vs w)	9.01	6.7e-10	0.293	0.227 - 0.360

Table 3: Statistics associated with grammatical and syntactical cues for expressing sentiment intensity. Differences in means were all statistically significant beyond the 0.001 level.

Sentiment datasets

- Test multiple domains: more believable!

3.4 Ground Truth in Multiple Domain Contexts

We next obtained gold standard (human-validated) ground truth regarding sentiment intensity on corpora representing four distinct domain contexts. For this purpose, we recruited 20 independent human raters from AMT (raters were all screened, trained, and data quality checked consistent with the process described in subsection 3.1.1 and Figure 2). All four sentiment-intensity annotated corpora are available for download from our website[14]:

1. Social media text: includes 4,000 tweets pulled from Twitter's public timeline (with varied times and days of posting), plus 200 contrived tweets that specifically test syntactical and grammatical conventions of conveying differences in sentiment intensity.
2. Movie reviews: 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.
3. Technical product reviews: 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.
4. Opinion news articles: includes 5,190 sentence-level snippets from 500 New York Times opinion editorials.

Comparison: other lexicons

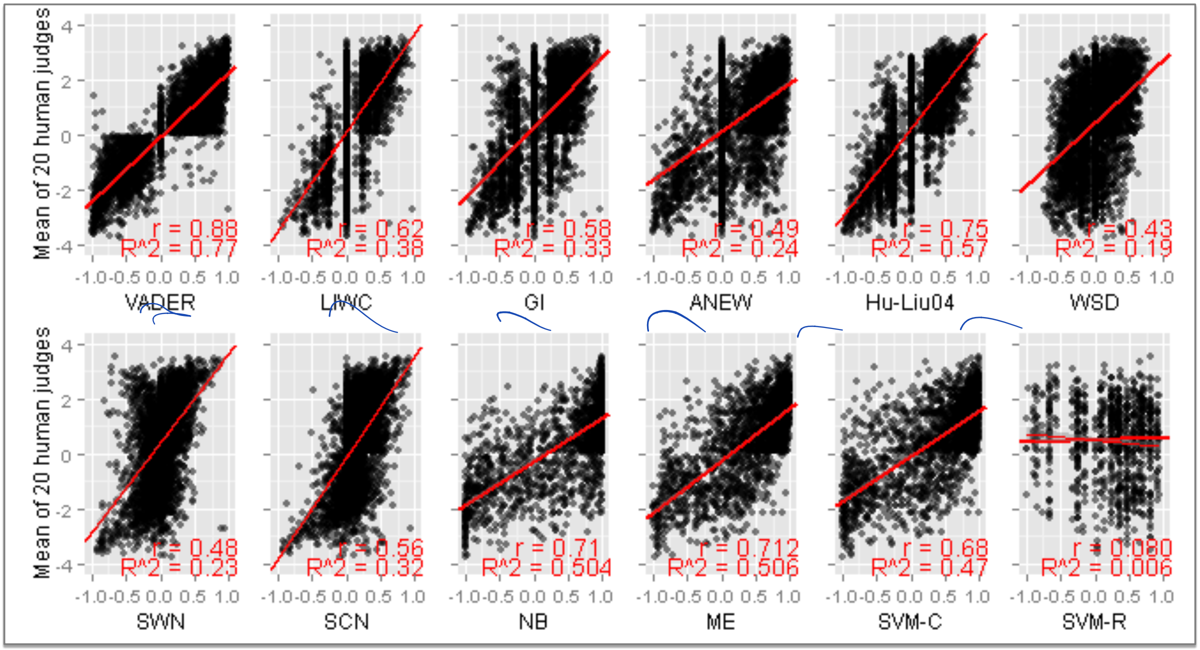

Figure 3: Sentiment scores from VADER and 11 other highly regarded sentiment analysis tools/techniques on a corpus of over 4K tweets. Although this figure specifically portrays correlation, it also helps to visually depict (and contrast) VADER's classification precision, recall, and F1 accuracy within this domain (see Table 4). Each subplot can be roughly considered as having four quadrants: true negatives (lower left), true positives (upper right), false negatives (upper left), and false positives (lower right).

Comparison: other lexicons

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score	Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	**2**	**1**	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	**1***	**2**	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	**3**	**3**	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	**4**	**7**	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	**5**	**5**	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	**6**	**4**	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	**7**	**9**	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	**8**	**8**	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	**9**	**6**	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	**1**	**1**	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	**2**	**2**	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	**3**	**3**	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	**7**	**7**	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	**5**	**5**	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	**4**	**4**	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	**9**	**9**	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	**8**	**8**	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	**6**	**6**	0.218	0.55	0.45	0.47

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

Comparison: machine learning

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

Table 5: Three-class accuracy (F1 scores) for each machine trained model (and the corpus it was trained on) as tested against every other domain context (SVM models for the movie and NYT data were too intensive for our multicore CPUs with 94GB RAM)

Same-domain ML
often beats
VADER

but
X-Domain ML
is worse!