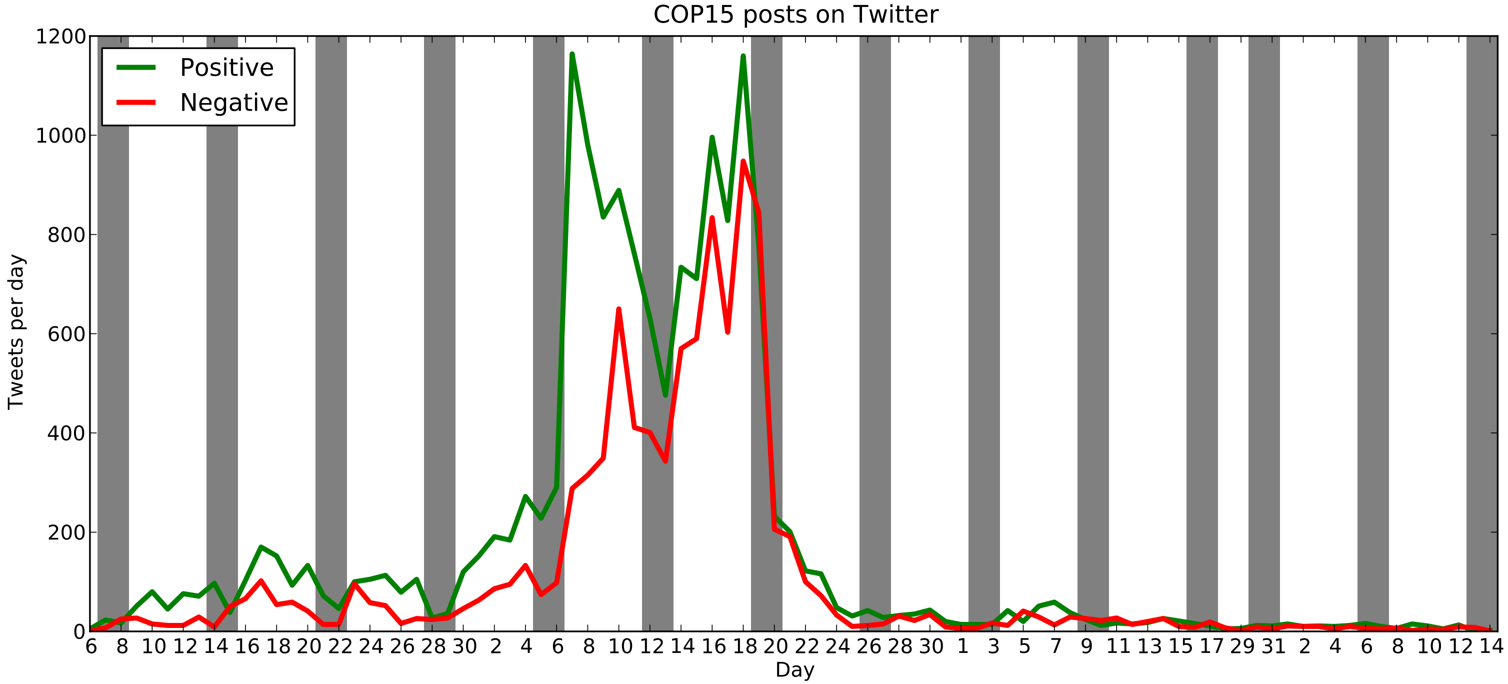
A new ANEW: Evaluation of a word list for sentiment analysis in microblogs

Finn Årup Nielsen

Department of Informatics and Mathematical Modelling
Technical University of Denmark

May 27, 2011

Sentiment analysis with word list



Since 2009 I have manually build a word list with valence for:

Temporal sentiment analysis on Twitter's COP15 posts in 2009/2010.

Retweet sentiment analysis (Hansen et al., 2011)

My word list

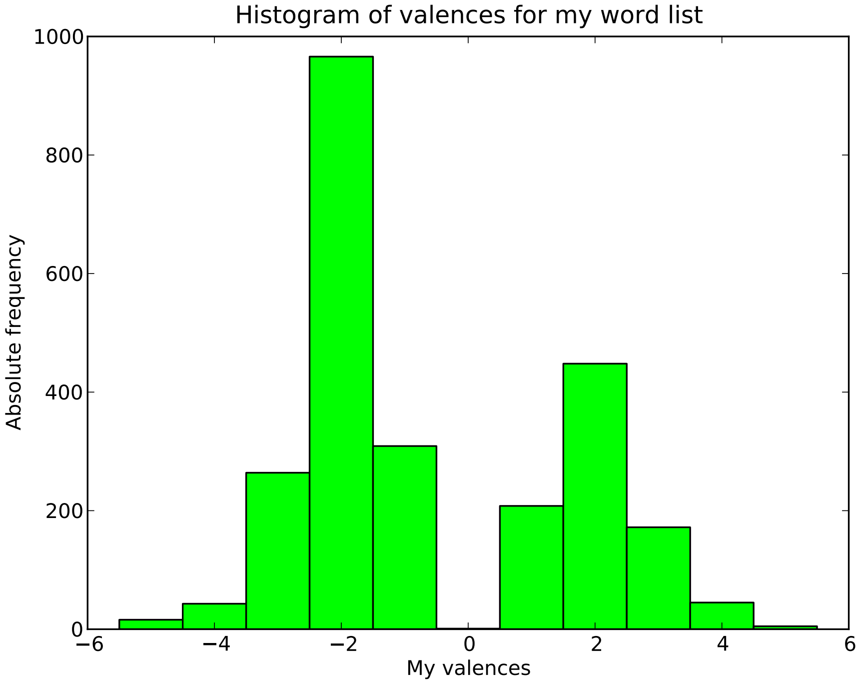


Figure 1: Histogram of my valences. Examples: abandon −2, abuse −3, ability +2, lol +3, greenwashing −3, hahaha +3, hurrah +5.

Each word scored between −5 (highly negative) and +5 (highly positive). Most words are negative, see histogram.

Latest version ("AFINN-111") has 2477 words.

Contains obscene words (Baudhuin, 1973; Sapolsky et al., 2008) and Internet slang (LOL, WTF, . . .)

Added words from Steve DeRose and Greg Siegle.

Available from homepage.

Example application on COP15 tweets

Low score: "I always get MAD furious and outraged by the stupid climate deniers' comments on every single news related to COP15 online. BLOODY HELL."

High score: "#cop15 Renaye - Our Planet : User comment : so cute! awesome wow amazing voice and great point keep on singing fantastic! http://ow.ly/HxeK"

Ambivalence: "Back home, BA wins luggage incompetence prize. Bag lost enroute to #cop15 was lost again on way home,plus 2 TV cases. Nice one Merry Xmas."

It seems to work reasonable.

But how well?

Wouldn't ANEW, a well-validated word list, be better?

Comparing word valence to ANEW

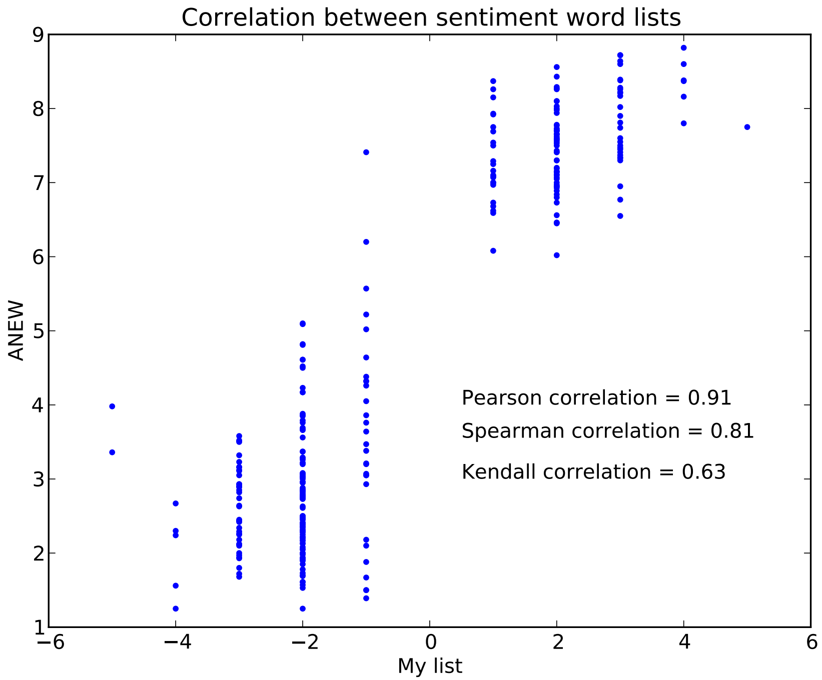


Figure 2: Correlation between ANEW and my new word list. Discrepancies: silly, hard, alert, mischief. Stemming issue: aggression/aggressive, alienation/alien, profit/profiteer.

ANEW (Affective Norms for English Words) (Bradley and Lang, 1999)

Compare the valence scores of each word from ANEW and my word list.

High correlation but the scoring of ANEW and my word list differ somewhat, see the scatterplot.

The correlation does not directly answer how well the word lists performs on sentiment analysis on microposts.

Alan Mislove AMT-labeled data

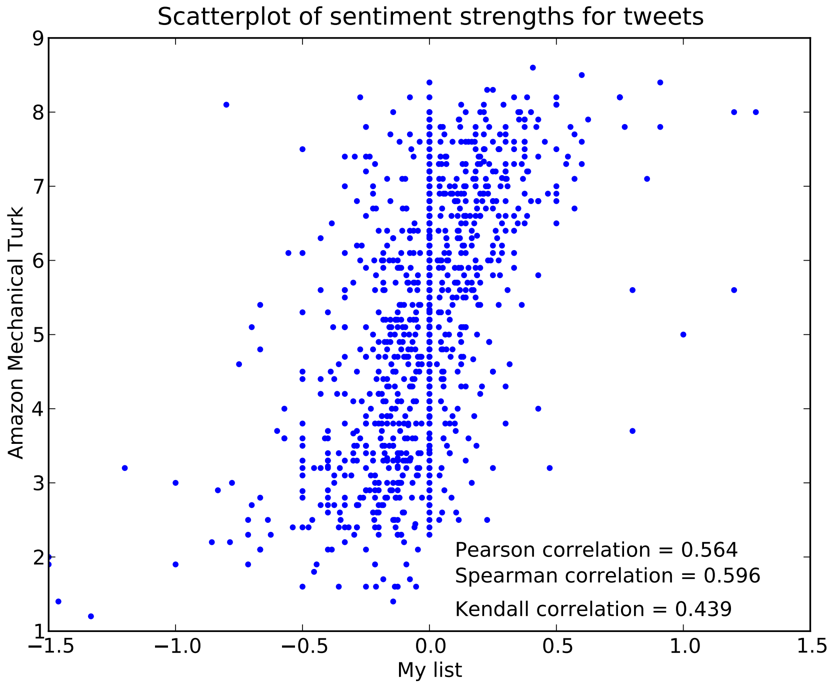


Figure 3: Scatter plot of sentiment strengths for 1,000 tweets with AMT sentiment plotted against sentiment found by application or my word list.

Alan Mislove data (Northeastern University, obtained through Sune Lehmann): 1'000 Amazon Mechanical Turk-labeled tweets. Each rated from 1 to 9 by 10 people.

Used in their "Twittermood" /" Pulse of the Nation" study (Biever, 2010).

Compare the score for 1'000 AMT-labeled texts: Scatter plot of My word list score against AMT mean score.

Example tweet scored with word lists

Also used other word lists from General Inquirer (GI) and OpinionFinder (OF) (Wilson et al., 2005) as well as the SentiStrength (SS) web-service (Thelwall et al., 2010)

GI: Polarity labeled, 3392 words used.

OF: Polarity labeled, 6442 words used.

Example with word scoring and tweet score:

	ear	infection	making	it	impossible	2	sleep	headed	2	the	doctors	2	get	new	prescription	so	f***ing	early	
My	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4	0	-4
AN	0	-3.34	0	0	0	0	2.2	0	0	0	0	0	0	0	0	0	0	0	-1.14
GI	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
OF	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
SS																			-2

Note: “infection”, “impossible”, “sleep”, “f***ing”.

AMT word list comparison

	My	ANEW	GI	OF	SS
AMT	.564	.525	.374	.458	.610
My		.696	.525	.675	.604
ANEW			.592	.624	.546
GI				.705	.474
OF					.512

Table 1: Pearson correlations between sentiment strength detections methods on 1,000 tweets. AMT: Amazon Mechanical Turk, GI: General Inquirer, OF: OpinionFinder, SS: SentiStrength.

Correlation matrix for sentiment strength detection.

SentiStrength has the highest correlation with ANEW.

My word list slightly ahead of ANEW, but no statistical test has been applied to answer whether this difference is important.

Sentiment analysis with word lists GI and OF I could not make perform well. I did not use extra information in these word lists.

An analysis with Spearman rank correlation gives qualitative similar results.

Word list size

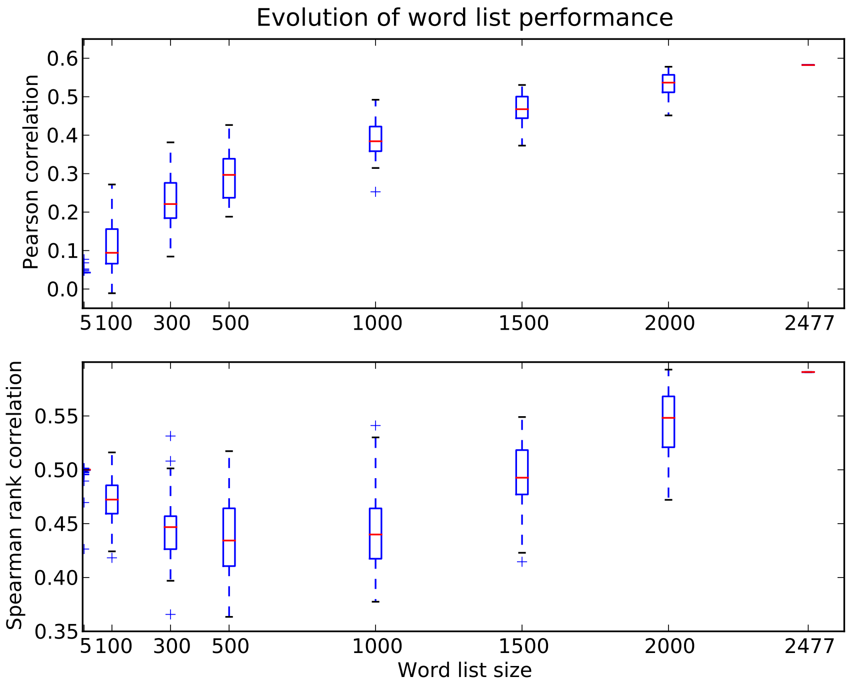


Figure 4: Performance growth with word list extension from 5 words 2477 words. Upper panel: Pearson, lower: Spearman rank correlation, generated from 50 resamples among the 2477 words.

Performance of my word list as the size is increased.

Still may be possible to increase the performance by adding more words.

But it may be more difficult to find words. The low-hanging fruit ("good", "bad") already taken.

Variants, ensembles, emoticons

Variant	Correlation
My word list (averaging scores, original)	0.564
My word list (averaging scores, other tokenization)	0.556
My word list (sum scores)	0.581
My word list (extreme valence)	0.543
ANEW variants	0.523–0.526
Ensemble word list	0.549
"Cheat" ensemble word list	0.580
My word list and emoticons	0.579
SentiStrength	0.610

Pearson correlation with AMT mean score.

Ensemble combines OF, GI, ANEW and my word list for average valence.

For "cheat" ensemble the word valences are determined from the word list that had the highest correlation with the AMT-labels.

Conclussions

My word list may be slightly ahead of ANEW for Twitter sentiment strength detection.

Word lists with sentiment strength for each word seem to be better than word lists with only polarity for sentiment strength detection.

The SentiStrength had the best performance. It has handling of negations, "booster words", misspellings, emoticons . . .

At a size of 2477 there are still many words missing. Performance may increase with addition of more words.

Valence score in my word list may be improved.

Word list is available.

References

Baudhuin, E. S. (1973). Obscene language and evaluative response: an empirical study. *Psychological Reports*, 32.

Biever, C. (2010). Twitter mood maps reveal emotional states of America. *The New Scientist*, 207(2771):14. DOI: 10.1016/S0262-4079(10)61833-7.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., and Etter, M. (2011). Good friends, bad news — affect and virality in Twitter. Accepted for The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011). http://arxiv.org/abs/1101.0510.

Sapolsky, B. S., Shafer, D. M., and Kaye, B. K. (2008). Rating offensive words in three television program contexts. BEA 2008, Research Division.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558. http://www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.doc.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.