An Empirical Evaluation of Sentiment Analysis on Movie Scripts

Andreas van Cranenburgh, University of Groningen
DH Benelux 2020

Sentiment analysis is a natural language processing method to analyze evaluative language (ranging from a single sentence to a whole review) as positive, negative, or neutral. The notion of sentiment is thus one dimensional, and usually discrete as well; it clearly ignores many nuances and complexities of emotions. Although a crude measure, insofar as it can be accurately predicted, it is a useful variable to analyze.

Sentiment analysis has also been applied in digital humanities (see Kim & Klinger 2019). However, this implies that the technique is applied to domains which it was never designed for. Sentiment analysis tools have been designed to classify explicit evaluative language such as reviews or opinions on social media. Whether it can capture more implicit emotions in other text types has not been adequately validated.

In particular, this is relevant for research that claims that the number of sentiment words in a moving window can be used to extract emotional arcs from novels (Jockers 2015; Reagan et al 2015). The resulting emotional arcs are purportedly instances of six universal plot shapes, a claim which has been criticized on various technical grounds (Swafford 2015; Enderle 2015).

We consider a more basic problem: is sentiment analysis of narrative texts reliable? We evaluate commonly used sentiment analysis methods on movie scripts obtained from www.imsdb.com. Jockers (2015) and Reagan et al (2015) apply sentiment analysis to longer pieces of text, but this is difficult to validate and annotate. Moreover, while aggregating sentiment scores of longer chunks of text may average out some of the errors, evaluating at the sentence level provides the most fine-grained picture of the performance. We therefore focus our validation effort on the task of sentence-level sentiment analysis. Students annotated 100 random sentences from 8 movie scripts (total 800 sentences with a minimum of 50 characters) with labels positive, negative, neutral. Two different sentiment analysis techniques are evaluated on this dataset:

1. LEX: A lexicon-based word counting approach using the Hu & Liu (2004) Opinion Lexicon; the count of negative words is subtracted from the count of positive words.
2. The rule-based VADER method (Hutto & Gilbert 2014), which has rules to deal with negation and other issues; a threshold is applied to convert the score to one of the three possible labels in the annotations: > 0.4 is positive and < -0.4 is negative.

See Table 1 for the results. For all but one of the movies, VADER outperforms LEX by a substantial margin, and is therefore the preferred method. We also find considerable variance between movies. In addition, the majority of the labels is neutral; when only negative or positive labels are evaluated (F1_neg and F1_pos), scores are lower. Table 2 compares the performance with other datasets reported by Ribeiro et al (2016), showing that our scores are in line with previous results.
In conclusion, the performance on narrative text is not substantially lower than with other domains, but the error rate is substantial. The scores reported here confirm that there is substantial cross-domain as well as in-domain variance. Where possible, annotating domain-specific data for training a machine learning system should be preferred over relying on off-the-shelf lexicon-based methods.

Movie	LEX	VADER		
	Acc %	Acc %	F1_neg	F1_pos
Die Hard	63	70	51	33
The Shining	78	84	18	60
Romeo & Juliet (1995)	42	58	36	15
Avengers	50	61	38	33
Inglourious Basterds	58	63	40	51
Double Indemnity	63	72	44	53
Inception	73	69	55	48
Alien (1979)	55	60	38	20
Average (mean)	60.3	67.6	40.0	39.1

Table 1: Three-label evaluation of sentence-level sentiment analysis of movie scripts.

Dataset	LEX	VADER
	Acc %	Acc %
Movies (this work)	60.3	67.6
Tweets SemEval	60.4	60.2
Tweets RND III	63.9	60.1
Comments BBC	55.0	49.4
Comments NYT	44.6	48.0

Table 2: Comparison with results reported in Ribeiro et al (2016).

References

Enderle, Scott (2015). A plot of Brownian noise. https://senderle.github.io/svd-noise

Hu, Minqing & Bing Liu (2004). "Mining and summarizing customer reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper)*, Seattle, Washington, USA, Aug 22-25.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109

Jockers, Matthew L. (2015). Revealing Sentiment and Plot Arcs with the Syuzhet Package. http://www.matthewjockers.net/2015/02/02/syuzhet/

Kim, Evgeny & Roman Klinger (2019). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. https://arxiv.org/abs/1808.03137

Swafford, Annie (2015). Continuing the Syuzhet discussion. https://annieswafford.wordpress.com/2015/03/07/continuingsyuzhet/

Reagan, A.J., Mitchell, L., Kiley, D. *et al.* (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **5,** 31. https://doi.org/10.1140/epjds/s13688-016-0093-1

Ribeiro, F.N., Araújo, M., Gonçalves, P. *et al.* (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5,** 23. https://doi.org/10.1140/epjds/s13688-016-0085-1