Annotation and Prediction of Movie Sentiment Arcs

Andreas van Cranenburgh, University of Groningen
Computational Stylistics Workshop
on Emotion and Sentiment Analysis in Literature, 2022

Some narratologists have argued that all stories derive from a limited set of archetypes. Specifically, Vonnegut (2005) claims in his Shapes of Stories lecture that if we graph the emotions in a story over time, the shape will be an instance of one of six basic story shapes. The work of Jockers (2015) and Reagan et al. (2016) purports to confirm this hypothesis empirically using automatic sentiment analysis (rather than manual annotations of story arcs) and algorithms to cluster story arcs into fundamental shapes. Later work has applied similar techniques to movies (Del Vecchio et al., 2019). This line of work has attracted criticism. Swafford (2015) argues that sentiment analysis needs to be validated on and adapted to narrative text. Enderle (2016) argues that the various methods to reduce story shapes to the putative six fundamental types are actually producing algorithmic artifacts, and that random sentiment arcs can also be clustered into six "fundamental" shapes.

In this paper I will not attempt to find *fundamental* (or even *universal*) story shapes, but I will take the observed story shape for each narrative as is, without trying to cluster them into archetypes. My aim is to perform an empirical validation of how well basic sentiment analysis tools can reproduce a sentiment arc obtained through manual annotation based on narrative text. Rather than considering novels as narratives, I consider movies, since the annotation of movies, when done in real time, is less time consuming. In a previous abstract, I considered the task of predicting the annotated sentiment of individual sentences from movie scripts (van Cranenburgh, 2020), and concluded that sentiment analysis tools achieve comparable performance on narrative text as compared to reviews and social media text (*pace* Swafford 2015). In this abstract I consider the task of predicting the overall sentiment as annotated based on watching the movie. This task is more challenging since the connection between the narrative sentiment and the narrative text is potentially more distant.

DH bachelor students selected 18 movies from various genres and annotated them with sentiment arcs. Our notion of sentiment is the same as Vonnegut's notion of good versus ill fortune at each point in the narrative. Each annotator watched one or more movies and annotated the overall sentiment for each minute of the movie with a label representing a negative, neutral or positive sentiment. These labels form the ground truth for our task.

The next step is to see how well simple sentiment analysis methods can approximate these annotations based on the subtitles of the movies. For sentiment analysis we use the VADER (Hutto & Gilbert, 2014) rule-based sentiment analysis tool for English, as part of the NLTK library. Given the annotations and subtitles of a movie, we can plot a sentiment arc by taking a moving average of the annotations or predicted sentiment of the preceding 10 minutes (i.e.,

each point in the plot will represent an aggregate sentiment score based on multiple annotations or subtitles). Figure 1 shows such a visualization of two movies. For the first movie, the shape of the sentiment arc is reasonably well predicted. For the second movie, there are some matching peaks and valleys, but the predicted arc is admittedly less accurate. Notice that in the second movie, there are gaps in the line for the sentiment based on subtitles, since these parts of the movie contain no dialogue.

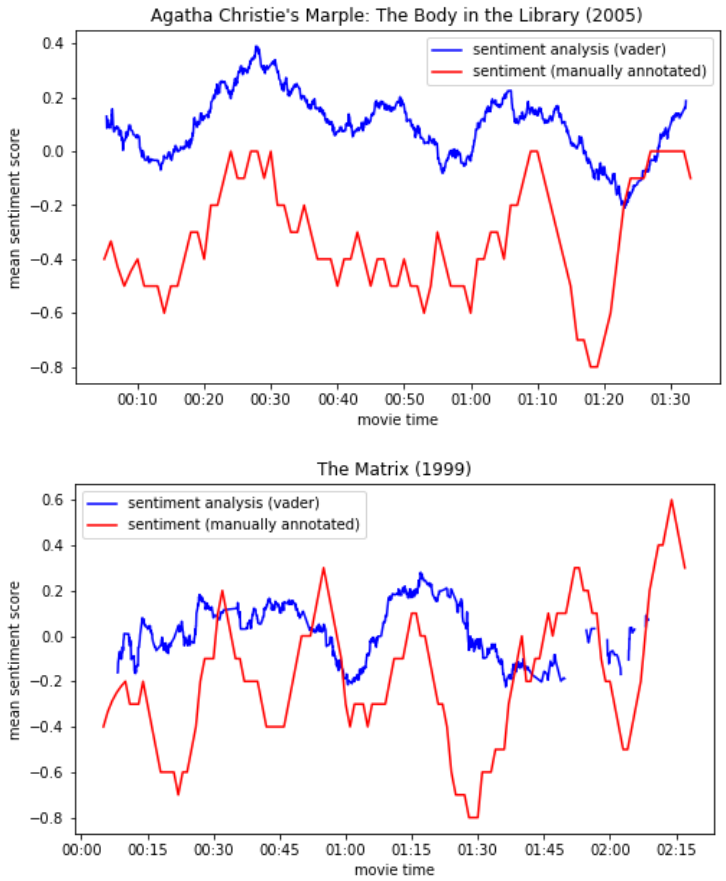

Figure 1: Sentiment arcs, automatically predicted vs annotated.

We can also try to predict the annotated labels using the subtitles of the movie as input. We group the subtitles into chunks of 1 minute, corresponding to the intervals that were annotated by the annotators. We apply VADER to all subtitle text within each chunk. VADER returns continuous scores in the range [-1, 1]. We discretize these scores using a threshold t. For each score s, if $s < -t$, the score is converted to a negative label, if $s > t$, the score is converted to a positive label; other scores are converted to neutral labels. See Table 1 for the classification scores. The table shows per-label micro F1-scores and the overall accuracy score.

	neg	neu	pos	acc
The Matrix (1999)	31.7	66.7	16.2	51.8
The Social Network (2010)	27.8	27.8	47.6	35.1
Manos (1966)	40.7	48.5	13.3	41.4
Agatha Christie's Marple: The Body in the Library (2005)	45.2	54.5	17.0	41.9
The Amazing Spider-man (2012)	6.0	46.0	28.6	31.2
Gravity (2013)	25.6	56.3	22.2	41.1
Batman v Superman: Dawn of Justice (2016)	21.4	52.3	10.3	37.7
The Notebook (2004)	23.8	38.6	48.1	40.2
Little Women (1994)	12.0	43.0	37.6	34.2
Spider-man: Homecoming (2017)	28.0	51.8	40.5	43.2
Psycho (1960)	21.4	59.3	39.1	45.5
The Hunger Games (2012)	7.1	48.9	38.4	37.3
Star Wars: The Force Awakens (2015)	19.0	58.5	23.5	43.2
Curse of the Black Pearl (2003)	34.4	10.8	22.0	22.4
Up (2009)	22.2	35.0	32.3	31.5
Spider-man (2002)	26.7	42.2	23.9	31.9
The Lion King (1994)	42.4	17.0	47.3	36.9
The Perks of Being a Wallflower (2012)	33.3	40.6	52.9	44.4
macro avg	26.0	44.3	31.2	38.4

Table 1: Sentiment classification scores (F1 scores for labels, accuracy for overall score); sentiment predicted using VADER with threshold=0.75.

VADER is a very simple lexicon- and rule-based sentiment analysis tool. Perhaps a machine learned sentiment analysis performs better? We train a regularized Logistic Regression model with tf-idf bag-of-words (BoW) features to predict the annotated labels. We report scores with 5-fold crossvalidation; see Table 2. The BoW model obtains a similar overall score compared to the VADER model.

	precision	recall	f1-score	support
neg	43.7	35.8	39.3	825
neu	39.1	49.1	43.6	768
pos	27.7	25.0	26.3	468
accuracy			38.3	2061

Table 2: Sentiment classification scores (cross-validated); sentiment predicted using Logistic Regression on tf-idf Bag-of-Words.

In conclusion, it appears that predicting sentiment arcs from narrative text is a challenging task. Subtitles of movies are not necessarily comparable to the text of novels as considered by Jockers (2015) and Reagan et al. (2016) or the movie scripts used by Del Vecchio et al. (2019), since subtitles only contain dialogue. Still, these results call for caution in relying on sentiment analysis to obtain sentiment arcs for narratives, echoing Swafford's (2015) critique.

This abstract has only applied very basic methods. More sophisticated machine learning methods such as word embeddings and transformers should be evaluated in future work to get a better upper bound on how well the sentiment arcs can be predicted from text. Another promising direction for future work is to replace the three-label classification approach with an evaluation that evaluates the shape of the plots, regardless of amplitude/intensity, and without the arbitrary discretization of sentiment scores. Finally, rather than a quantitative evaluation of predictability, a qualitative study of the sentiment arcs would undoubtedly provide many insights.

References

Del Vecchio, Marco, Kharlamov, A., Parry, G., & Pogrebna, G. (2021). Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries. *Journal of the Operational Research Society*, *72*(5), 1110-1137. https://doi.org/10.1080/01605682.2019.1705194

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

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/

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

Reagan, Andrew J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science, 5(1), 1-12. https://doi.org/10.1140/epjds/s13688-016-0093-1

van Cranenburgh, Andreas (2020). An Empirical Evaluation of Sentiment Analysis on Movie Scripts. DH Benelux 2020. https://zenodo.org/record/3862158

Vonnegut, Kurt (2005). At the blackboard: Kurt Vonnegut diagrams the shapes of stories. https://www.laphamsquarterly.org/arts-letters/blackboard