An Empirical Evaluation of Sentiment Analysis on Movie Scripts

Andreas van Cranenburgh,
University of Groningen

June 4, 2020, DHBenelux

@andreasvc

Definition

Sentiment analysis: automatically identify positive and negative language

Definition

Sentiment analysis: automatically identify positive and negative language

- Tools based on word lists and machine learning
- Designed to analyze product reviews and social media posts (i.e., evaluative language)

Definition

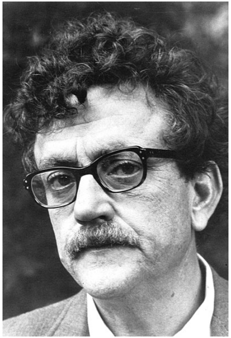
Sentiment analysis: automatically identify positive and negative language

- Tools based on word lists and machine learning
- Designed to analyze product reviews and social media posts (i.e., evaluative language)

Our research question:
How well does sentiment analysis work on narrative texts?

Kurt Vonnegut: The simple shapes of stories

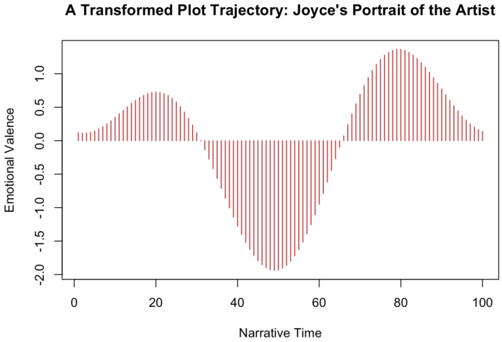
https://youtu.be/oP3c1h8v2ZQ?t=20



The shapes of stories:
a rejected master thesis topic ...

More information:
https://www.brainpickings.org/2012/11/26/kurt-vonnegut-on-the-shapes-of-stories/

Automatic story shape detection?



- Estimate emotions: count sentiment words
- Chop text into chunks of *x* words, count sentiment in each chunk
- "Discover" plot shapes with math tricks:
  - Fourier transform
  - Principal Components (SVD)
  - Hierarchical clustering
  - Self-organizing map

Seems to confirm that all novels have a few basic story shapes!

Jockers (2015) `http://www.matthewjockers.net/2015/02/02/syuzhet/`
Reagan et al (2015) `https://doi.org/10.1140/epjds/s13688-016-0093-1`

The problem: signal vs noise

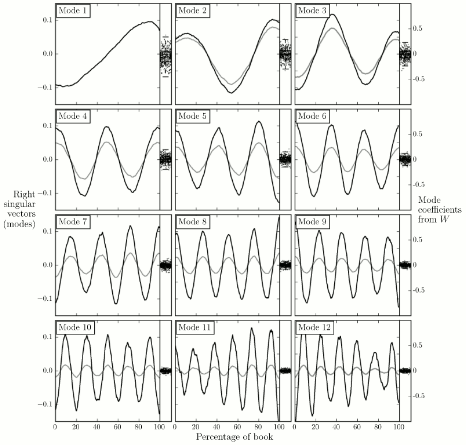


Figure 3 Top 12 modes from the singular value decomposition of 1,327 Project Gutenberg books. We

Critique by Enderle (2015):

- These plots just show sine waves with an increasing number of peaks
- Enderle argues that SVD is just modeling random noise, not any sentiment signal
- A completely randomly generated dataset also produces these plots!

Scott Enderle (2015). `https://senderle.github.io/svd-noise/`

Swafford's critique of sentiment analysis

"All approaches—from the lexicon-based approaches to the more advanced Stanford parser–have difficulty with anything that doesn't sound like a tweet or product review, which is not surprising."

- Word counting using lexicons is an extremely naive method (ignores context).
- Stanford parser: 80-85% accuracy on sentiment analysis . . . of movie reviews
- Literary texts much more ambiguous and nuanced.
- NLP methods are usually very domain dependent (i.e., they work well on the kind of data they were trained on).

https://annieswafford.wordpress.com/2015/03/07/continuingsyuzhet/

Swafford's critique of sentiment analysis

"All approaches—from the lexicon-based approaches to the more advanced Stanford parser–have difficulty with anything that doesn't sound like a tweet or product review, which is not surprising."

- Word counting using lexicons is an extremely naive method (ignores context).
- Stanford parser: 80-85% accuracy on sentiment analysis ...of movie reviews
- Literary texts much more ambiguous and nuanced.
- NLP methods are usually very domain dependent (i.e., they work well on the kind of data they were trained on).

Takeaway: we need annotated data to benchmark and train sentiment analysis systems on books (or movies etc).

https://annieswafford.wordpress.com/2015/03/07/continuingsyuzhet/

Approach

Annotation:

- Download 8 movie scripts from `www.imsdb.com`
- Pick 100 random sentences with $>$ 50 characters
- Annotate sentences with label positive, negative, neutral

Approach

Annotation:

- Download 8 movie scripts from `www.imsdb.com`
- Pick 100 random sentences with $>$ 50 characters
- Annotate sentences with label positive, negative, neutral

Evaluate:

- LEX: Opinion Lexicon (Hu & Liu 2004)
- VADER (Hutto & Gilbert 2014); as implemented in NLTK
  Convert scores to labels with threshold at -0.4 and 0.4.

Results

Three-label accuracy scores:

Movie	LEX Acc %	VADER Acc %
Romeo & Juliet (1995)	42	58
Alien (1979)	55	60
Avengers	50	61
Inglourious Basterds	58	63
Inception	73	69
Die Hard	63	70
Double Indemnity	63	72
The Shining	78	84
Average (mean)	60.3	67.6

- high (in-domain) variance
- VADER much better than LEX

Results

Most labels are neutral; what is performance on other labels?

Movie	VADER		
	Acc %	F1 neg	F1 pos
Romeo & Juliet (1995)	58	36	15
Alien (1979)	60	38	20
Avengers	61	38	33
Inglourious Basterds	63	40	51
Inception	69	55	48
Die Hard	70	51	33
Double Indemnity	72	44	53
The Shining	84	18	60
Average (mean)	67.6	40.0	39.1

- Low performance on positive and negative labels

Comparison with other datasets

Sentence-based classification of movie reviews (SST-2)
with latest deep learning methods:

Model	Acc %
XLNet-Large (ensemble) (Yang et al., 2019)	96.8
MT-DNN-ensemble (Liu et al., 2019)	96.5
...	

http://nlpprogress.com/english/sentiment_analysis.html

Comparison with other datasets

Sentence-based classification of movie reviews (SST-2)
with latest deep learning methods:

Model	Acc %
XLNet-Large (ensemble) (Yang et al., 2019)	96.8
MT-DNN-ensemble (Liu et al., 2019)	96.5
...	

However: binary classification! (positive, negative)
not comparable to three label classification;
e.g., random guessing gives 50% vs 33%

http://nlpprogress.com/english/sentiment_analysis.html

Comparison with other datasets

Dataset	LEX Acc %	VADER 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

- high (cross-domain) variance
- higher performance on movie scripts than tweets/reviews!

Ribeiro et al (2016). SentiBench: a benchmark comparison of state-of-the-practice sentiment analysis methods. https://doi.org/10.1140/epjds/s13688-016-0085-1

Conclusions

Takeaways:

- Don't compare classification results with 2 vs 3 labels!
- high cross- and in-domain variance
- VADER much better than LEX
- low performance on positive and negative labels

Answer to research question:
higher performance on movie scripts than tweets/reviews!

Conclusions

Takeaways:

- Don't compare classification results with 2 vs 3 labels!
- high cross- and in-domain variance
- VADER much better than LEX
- low performance on positive and negative labels

Answer to research question:
higher performance on movie scripts than tweets/reviews!

Open questions:

- Inter annotator agreement
- Effect of aggregating sentiment scores
- How much room for improvement with sufficient data and deep learning?
- What amount of errors is acceptable?

Thanks to my students for the annotation work!