18 January 2025

Ruining Emotional Arcs: Robustness of Sentiment Analysis to Textual Noise

Artem Suslov
Hokkaido University, Faculty of Humanities and Human Sciences, Laboratory of Western Literature
artem.suslov.b0@elms.hokudai.ac.jp

This work was supported by JST SPRING, Grant Number JPMJSP2119

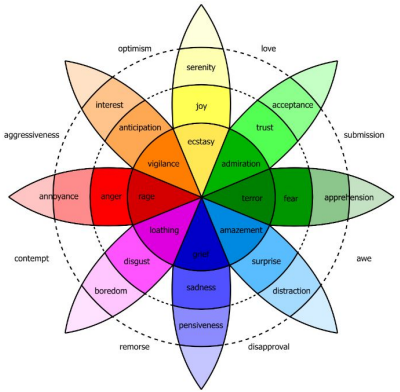
Outline

1. Sentiment Analysis in Computational Literary Studies
2. Emotional Arcs
3. DH for Japanese: few corpora and OCR quality
4. DH for Japanese: Oseti vs. BERT-based models
5. Noise Imitation
6. Sentiment Deviation
7. Arcs Distortion

Sentiment Analysis

1. Sentiment Analysis (SA) is a broad term for instruments that analyze “people's opinions, sentiments, appraisals, attitudes, and emotions towards entities and their attributes (Liu 2020).”

Categorical:



Plitchuk’s wheel of emotions (taken from Kim&Klinger)

SA Output

Polarity scores:

Sentence sentiment score = s{n}:
-1 <= s{n} <= +1
E.g.,
s{1} = 0;
s{2} = 0.34;
s{3} = -0.7;
etc.

SA in Computational Literary Studies: Emotional Arcs

M. Jockers suggested to smoothen raw sentiment data to build sentiment arcs and track plot development.

(Jockers and Thalken 165)

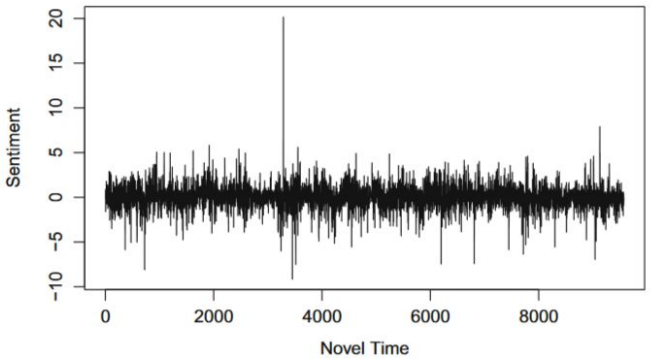


Fig. 14.1 Raw sentiment values in *Moby Dick*

Reagan et al. classified sentiment arcs into 6 typical patterns and attempted genre classification.

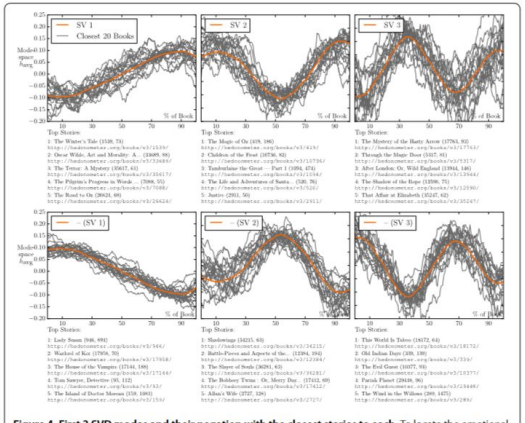


Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W. In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

(Reagan et al. 7)

SA in Computational Literary Studies: Emotional Arcs

M. Jockers suggested to smoothen raw sentiment data to build sentiment arcs and track plot development.

(Jockers and Thalken 170)

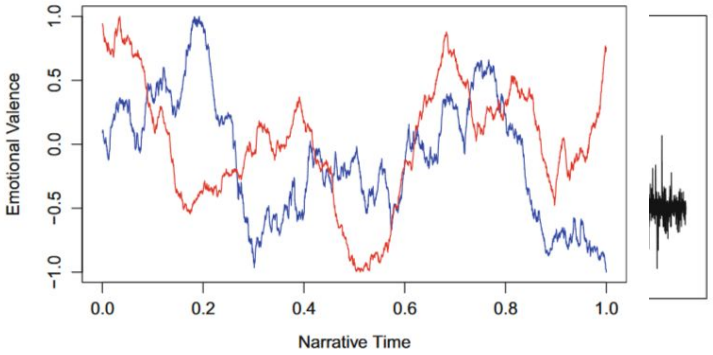


Fig. 14.6 Moby Dick and sense and sensibility with rolling means

Reagan et al. classified sentiment arcs into 6 typical patterns and attempted genre classification.

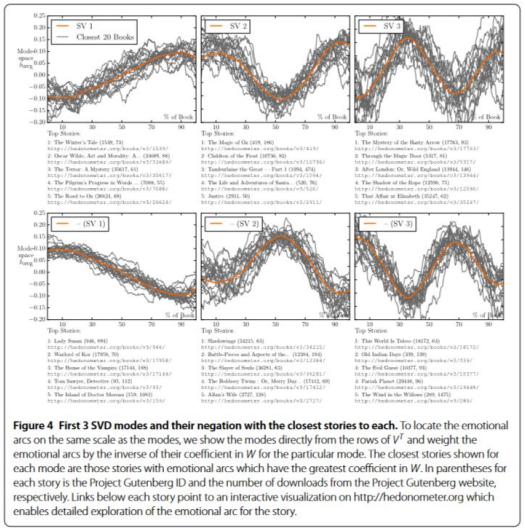


Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W. In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

(Reagan et al. 7)

Problem 1. Optical Character Recognition (OCR) Quality

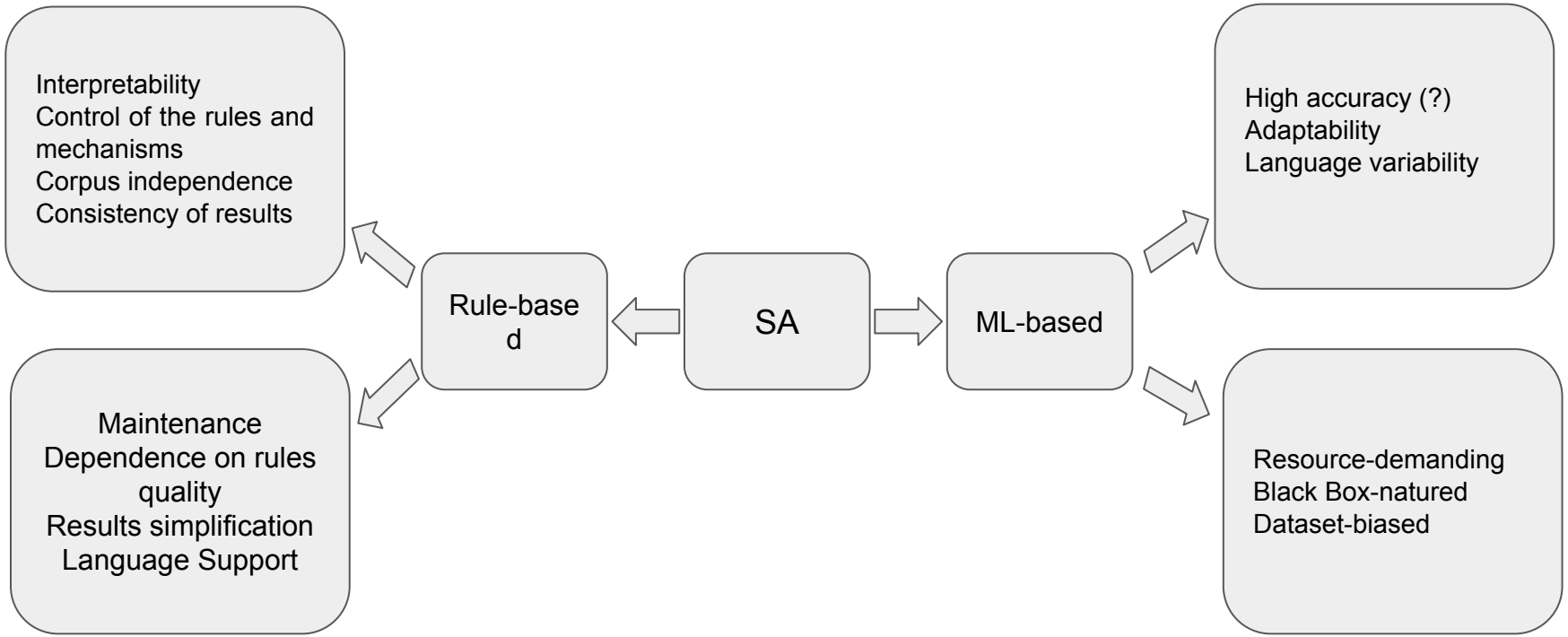
As any underrepresented in DH language, Japanese language lacks machine-ready, verified corpora of literary texts (e.g. Project Gutenberg, Aozora Bunko).

The quality of OCR for Japanese texts significantly varies:

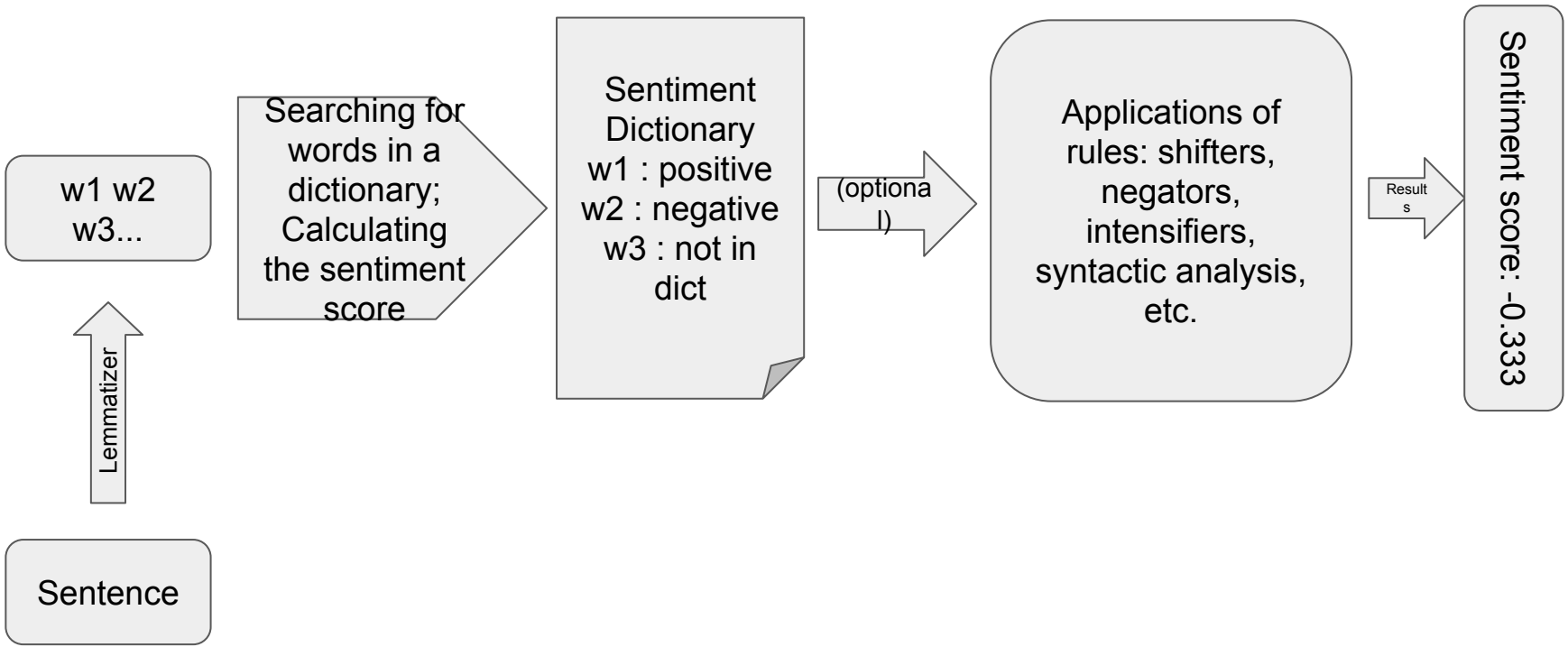
1. For classical texts:
   - Optimistic but reachable expected quality is 80% (Yamamoto and Osawa)
2. For vouchers:
   - Japanese OCR : 24.6%;
   - Google Vision : 42.6%;
   - Fast Accounting Robota : 97% (Fujitake).

For common modern printed texts, commercial solutions as Abbyy FineReader provide satisfactory results but still not free from OCR-errors.

Problem 2. Models and their Performance



Rule-Based SA Algorithm



Oseti & VADER Characteristics

VADER (Valence Aware Dictionary and sEntiment Reasoner)
by (Hutto & Gilbert 2014)
Lexicon and Rule-based
Language: English
provides four scores:

Positive: Proportion of text that is positive.

Negative: Proportion of text that is negative.

Neutral: Proportion of text that is neutral.

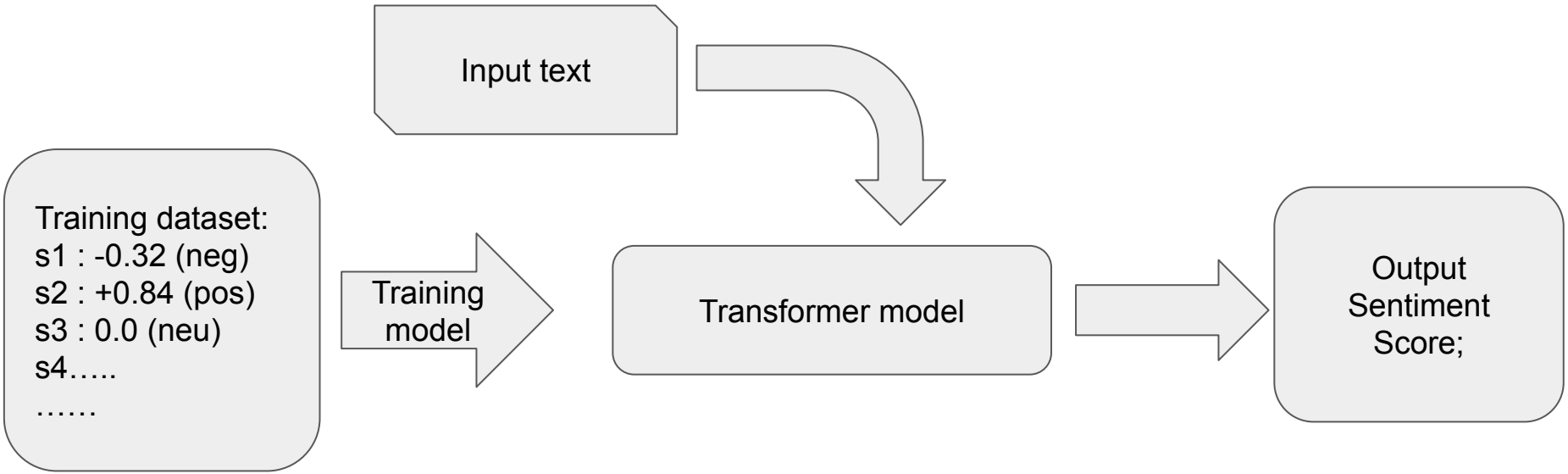
Compound: A normalized, weighted composite score that ranges from -1 (extremely negative) to +1 (extremely positive).

Oseti
by (Ikegami 2021)
Lexicon-based (considers negation)
Language: Japanese
Source of lexicon: 日本語評価極性辞書 (Dictionary of Japanese Language Evaluation)
Outputs the score ([-1, +1]) which is calculated as weighted difference between a number of negative- and positive-marked words in the sentence.

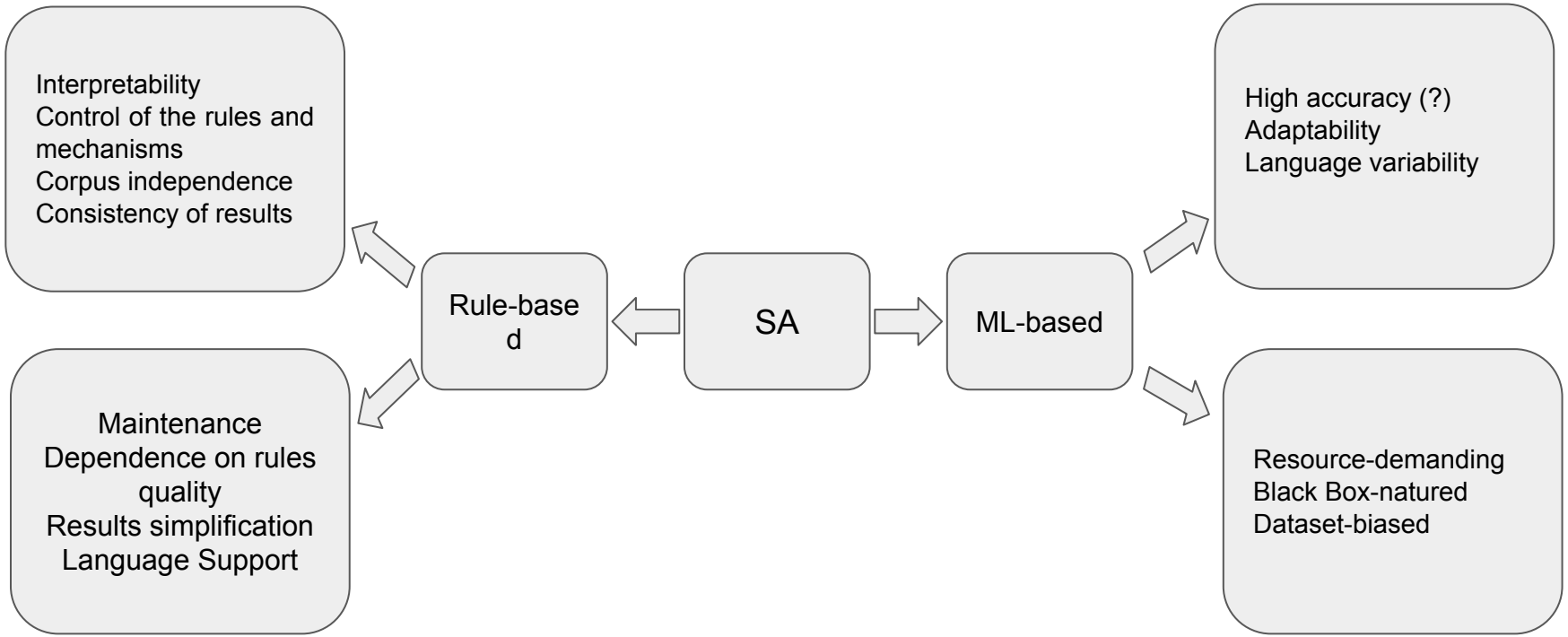
Dictionary-based Oseti's Pros and Cons.

1. Benefits:
   - Clear algorithm;
   - Better interpretability;
   - Consistency of results;
   - Good running speed (even in the interpreted language).
2. Problems:
   - Oversimplification of internal textual sentiment;
   - Lack of support (dependencies issues).

ML-Based SA Algorithm



Problem 2. Models and their Performance



Why not transformers?

21 models on Hugging Face

Unappropriate training datasets both by the size and the thematic field:

- **bert-finetuned-japanese-sentiment (Phu 2023)**: 20000 Amazon review;
- japanese-sentiment-analysis(Patel 2022): 220 corporate financial reports (6,119 sentences);
- Finance-sentiment-ja-base (bards.ai 2023): Japanese financial news (about 5,000 sentences/phrases);
- Japanese Stock Comment Sentiment Model (c299m 2023): comments and discussions related to Japanese stocks (dataset not described);
- Others - undocumented.

Major Problem

While working with “noisy” corpora, are complicated ML-based SA models worth to be used instead of simpler dictionary- and rule-based models?

How does the increasing level of textual noise distort the outputs of these models?

Textual Noise Imitation

Corpus Preparation

51 text from the “Japanese Atomic Bomb Literature” anthology, among them:
short (fewer than 1000 s.) -16;
Medium (1000-3000 s.) - 9;
Long (over 3000 s.) - 26.

Noise Imitation

By each iteration 1000 characters were replaced with noise until the texts consists of 100% noise.
For every iteration, SA score were collected.

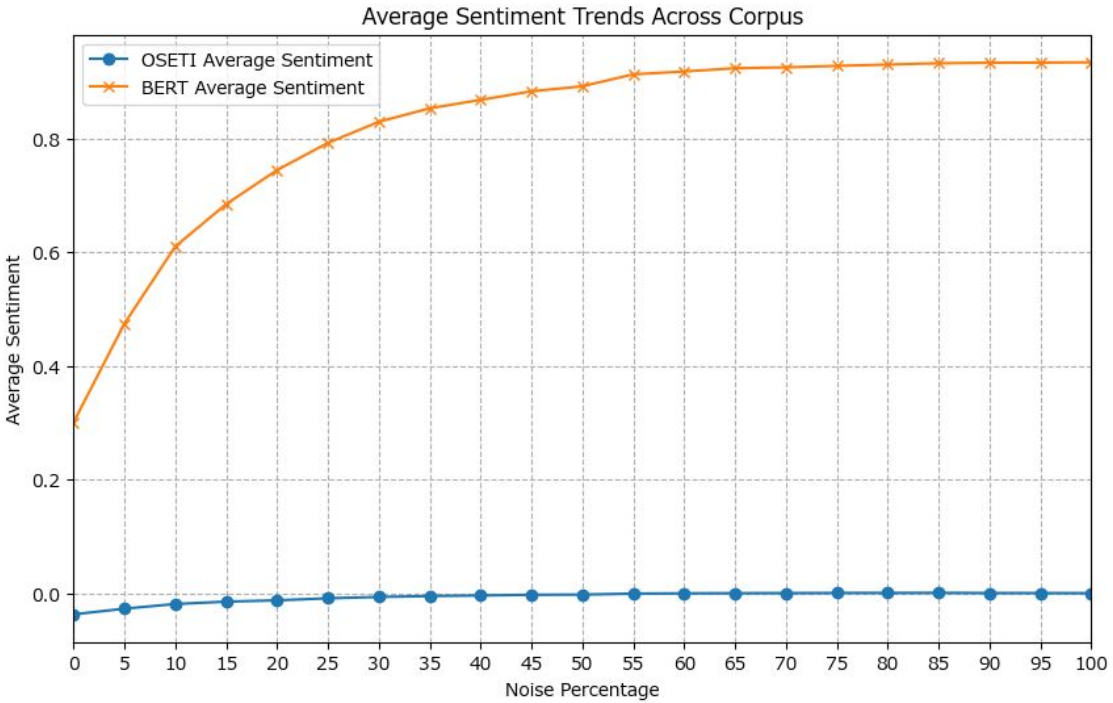
Total Sentiment Deviation Evaluation

The results for BERT and Oseti were grouped in 5% intervals. The original score was taken as 1, and then de facto normalized absolute error was calculated

Emotional Arcs Deviation Estimation

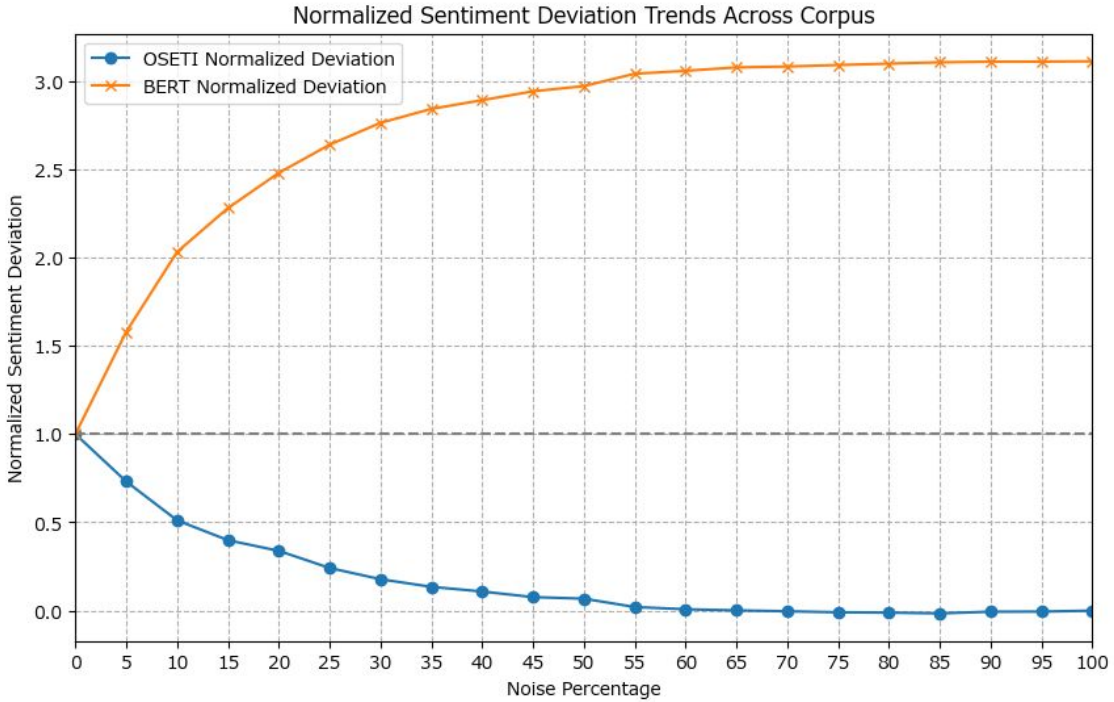
On the same intervals, sentiment arcs were graphed with manual adjustment of moving average window size. The similarity of emotional arcs shapes was visually estimated.

Total Sentiment Deviation - 1



BERT SA scores demonstrated a wider scope of response, while Oseti's ones were closer to neutral levels.

Total Sentiment Deviation - 2



1. Oseti's reaction to increasing noise level was more natural, while BERT behavior even under insignificant levels of noise was inadequate.

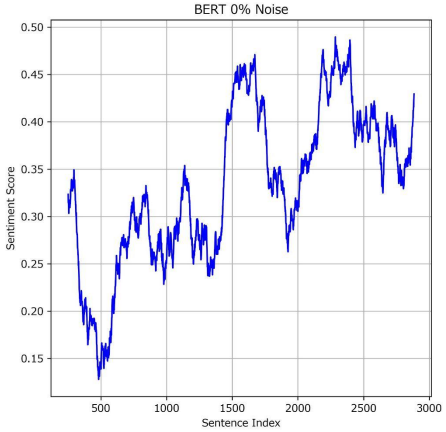
Emotional Arcs Resistance to Noise

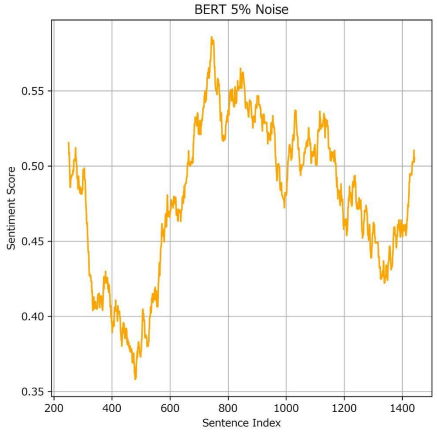
Major principles of sentiment arc distortion estimations:

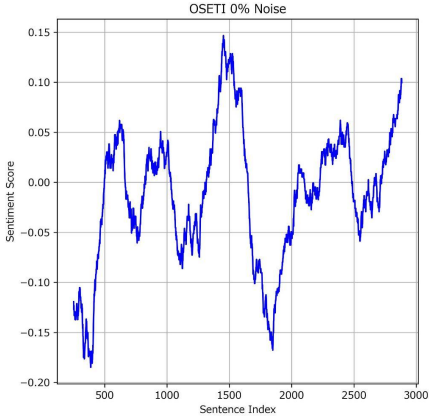
1. Does the shape trend remain the same?
2. Does the local minimums and maximums remain the same?
3. If being interpreted independently, would the interpretation be the same (position of culmination points, plot growth and decline, etc.)?

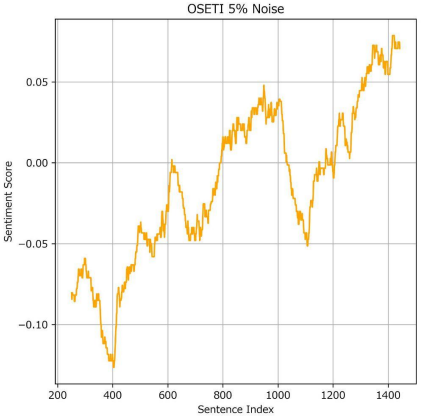
There could be a pure formal implication to estimate the change of sentiment arcs shapes (e.g., their approximations of the initial sentiment arcs to functions of particular types and then estimating the error of the approximations on different noise level from the initial one). However, the variation of sentiments shapes, difficulties with fixing moving average size put obstacles to such sort of formalizations for me.

Example 1. Oe Kenzaburo's *Hiroshima Notes 1*

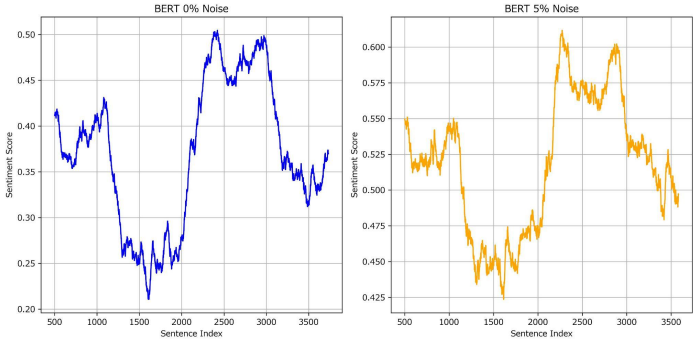


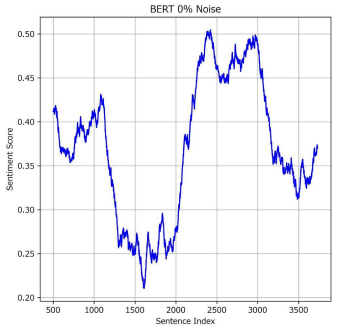


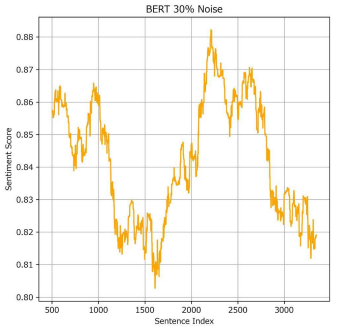


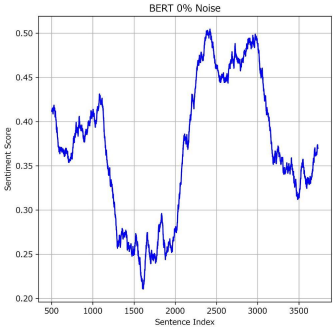


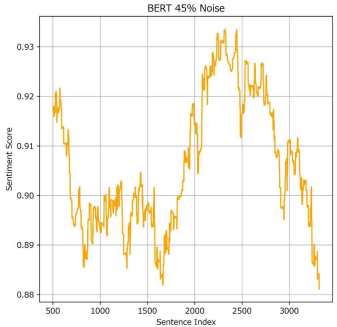
Example 2. Ida Momo's *American Hero (Part 7)*











Long Texts (over 3000 sentences) - 1

Title	BERT resistance	OSETI ~	OSETI outperforms
いいだ もも - アメリカの英雄1	20	35	yes
いいだ もも - アメリカの英雄2	20	10	no
いいだ もも - アメリカの英雄3	30	40	yes
いいだ もも - アメリカの英雄4	15	30	yes
いいだ もも - アメリカの英雄5	30	5	no
いいだ もも - アメリカの英雄6	25	5	no
いいだ もも - アメリカの英雄7	30	10	no
中山 士朗 - 死の影	15	20	yes
中本 たか子 - 死の鞭と光	5	5	equal

Long Texts (over 3000 sentences) - 2

Title	BERT resistance	OSETI ~	OSETI outperforms
亀沢 深雪 - 広島巡礼	0	0	equal
井上 光晴 - 地の群れ	30	35	yes
佃 実夫 - 赤と黒の喪章	15	25	yes
佐多 稲子 - 樹影	25	25	equal
原 民喜 - 夏の花	0	0	equal
大田 洋子 - 屍の街	25	20	no
安部 和枝 - 小さき十字架を負いて	5	5	equal
小久保 均 - 夏の刻印	15	50	yes
小田 勝造 - 同窓会は夏に	0	0	equal

Long Texts (over 3000 sentences) - 3

Title	BERT resistance	OSETI ~	OSETI outperforms
廣中 俊雄 - 炎の日	10	5	no
文沢隆一 - 重い車	5	15	yes
有吉 佐和子 - 祈禱	5	5	equal
栗田 藤平 - 青銅色の闇	5	5	equal
梶山 季之 - 実験都市	0	0	equal
武田 泰淳 - 第一のボタン	20	60	yes
生口 十朗 - 死者への勲章	15	10	no

Long Texts (over 3000 sentences) - 4

	BERT resistance	OSETI ~
average	15.38	17.31
median	15	10

Oseti outperforms in 9/26 cases
BERT outperforms in 8/26 cases
Both models demonstrated the same results in 9/26 cases

Medium Texts (1000-3000 sentences) - 1

Title	BERT resistance	OSETI ~	OSETI outperforms
中井 正文 - 名前のない男	5	5	equal
古浦 千穂子 - 風化の底	0	15	yes
夏堀 正元 - 聖地の女	10	15	yes
大江 健三郎 - ヒロシマ・ノ―ト 1	10	15	yes
岩崎 清一郎 - 過ぐる夏に	0	5	yes
川上宗薫 - 残存者	0	10	yes
斎木寿夫 - 死者は裁かない	5	15	yes
林 京子 - 同期会	0	5	yes
石田 耕治 - 雲の記憶	0	0	equal

Medium Texts (1000-3000 sentences) - 2

	BERT resistance, %	OSETI ~	OSETI outperforms
average	3.33	9.44	average
median	0	10	median

Oseti outperforms in 7/9 cases
BERT outperforms in 0/9 cases
Both models demonstrated the same results in 2/9 cases

Short Texts (< 1000 sentences) - 1

Title	BERT resistance	OSETI ~	OSETI outperforms
中里 喜昭 - 黄葵	0	15	yes
井上 光晴 - 「七〇年夏」への告発	5	5	equal
井上 光晴 - 「前科三犯」の被爆者	50	50	equal
井上 光晴 - プルトニウムの秋	10	15	yes
井上 光晴 - 原子力潜水艦をむかえる基地市民の 感覚	5	5	equal
井上 光晴 - 夏の客	5	15	yes
井上 光晴 - 手の家	5	10	yes
大江 健三郎 - ヒロシマ・ノ―ト 2	5	10	yes
大江 健三郎 - ヒロシマ・ノ―ト 3	15	10	no

Short Texts (< 1000 sentences) - 2

Title	BERT resistance	OSETI ~	OSETI outperforms
大江 健三郎 - 核状況のカナリア理論	15	20	yes
峡 草夫 - どくだみ草	0	20	yes
桂 芳久 - 氷牡丹	5	5	equal
桂 芳久 - 火と碑	5	15	yes
橋岡 武 - 八月二十三日の事	20	10	no
田 洋子 - 私と「原爆症」とについて	15	5	no
稲田 美穂子 - 見知られぬ旅	15	10	no

Short Texts (< 1000 sentences) - 3

Title	BERT resistance	OSETI ~
average	10.9375	13.75
median	5	10

Oseti outperforms in 8/16 cases
BERT outperforms in 4/16 cases
Both models demonstrated the same results in 4/16 cases

Conclusions

1. The distortion patterns of overall text sentiment significantly differ: while OSETI just ignores distorted by noise tokens, assigns them 0 sentiment, i.e. the total score tends to zero, BERT tends not to ignore these words and assigns them more positive scores.
2. Dealing with long texts, two models demonstrate compatible results.
3. On the corpus of medium length texts, Oseti significantly outperformed BERT, while BERT demonstrated almost complete inoperationability.
4. On the corpus of short texts, Oseti also outperforms BERT model.
5. In general, the dictionary-based model demonstrated a better performance on medium length and shorter texts, even though the model input in both cases was a tokenized sentence.

Conclusions

6. Overall, long texts seemed to demonstrate the sustainability up to 15% of noise; medium texts processed by Oseti may sustain up to 10% of noise; and short texts are sustainable up to 10% of noise for both models.

7. For CLS tasks, in which SA outcomes are processed in forms of aggregated sentiment scores and emotional arcs, existing transformers models, fine-tuned on limited and irrelevant corpora, do not provide any improvement in performance.

References - 2

Kim, Evgeny, and Roman Klinger. *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies*. 2019.

Liu, Bing (2020). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

Patel, Jignesh (2022). japanese-sentiment-analysis. In Hugging Face repository. Hugging Face. https://huggingface.co/jarvisx17/japanese-sentiment-analysis.

Phu, Christian (2023). bert-finetuned-japanese-sentiment. In Hugging Face repository. Hugging Face. https://huggingface.co/christian-phu/bert-finetuned-japanese-sentiment.

Reagan, Andrew J., et al. “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes.” *EPJ Data Science*, vol. 5, no. 1, Dec. 2016, p. 31. *DOI.org (Crossref)*, https://doi.org/10.1140/epjds/s13688-016-0093-1.

Yamamoto, Sumiko, and Tomejiro Osawa. “Labor saving for reprinting Japanese rare classical books : The development of the new method for OCR technology including kana and kanji characters in cursive style.” Journal of Information Processing and Management, vol. 58, no. 11, pp. 819-827.

References -1

Bards.ai (2023). Finance Sentiment JA (base). In Hugging Face repository. Hugging Face. https://huggingface.co/bardsai/finance-sentiment-ja-base.

c299m (2023). Japanese Stock Comment Sentiment Model. In Hugging Face repository. Hugging Face. https://huggingface.co/c299m/japanese_stock_sentiment.

Fujitake, Masato. *JaPOC: Japanese Post-OCR Correction Benchmark Using Vouchers*. arXiv:2409.19948, arXiv, 30 Sept. 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2409.19948.

Hutto, Clayton and Eric Gilbert (2014). “Vader: A parsimonious rule-based model for sentimentanalysis of social media text”. In: Proceedings of the international AAAI conference on web and social media. Vol. 8. 1, pp. 216–225.

Ikegami, Yukino. (2021). Oseti. In GitHub repository. GitHub. https://github.com/ikegami-yukino/oseti

Jockers, Matthew L., and Rosamond Thalken. *Text Analysis with R: For Students of Literature*. Springer International Publishing, 2020. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-39643-5.