Digital Humanities: Text-as-Data

Week 4 – Sentiment Analysis: Concepts, Techniques, and Cautions

Steven Denney
Leiden University

BA3 Korean Studies
November 07 2025

Comparative Political Studies
OnlineFirst

https://doi.org/10.1177/00104140251392537

Sage Journals

Original Article

Narrating the Nation: Colonial Memory and Citizen Attitudes in East and Southeast Asia

Dean Dulay [1] **and Jiyoung Ko** [2]

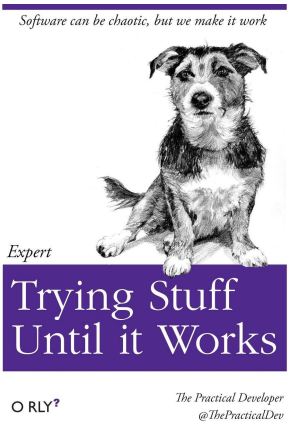
Abstract

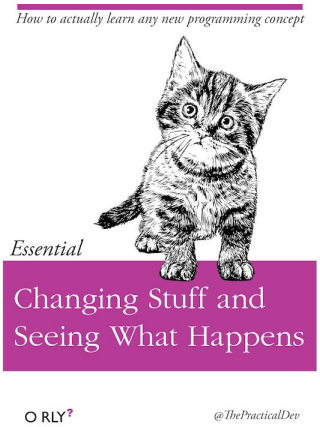
The postcolonial world exhibits stark variation in how citizens of formerly colonized states view their former colonizers. Some harbor continuing animosity, while others view former colonizer countries favorably. We argue that this variation can be explained by the type of national narrative postcolonial states construct upon independence. We conduct a comparative historical analysis of three formerly colonized Asian countries—the Philippines under the United States, Korea under the Japanese, and Indonesia under the Dutch—and find that distinct founding national narratives lead to differing contemporary public opinion. The continuity narrative in the Philippines, emphasizing positive aspects of American rule, has led to positive public opinion towards the United States. Korea's restoration narrative highlights an ancestral, precolonial identity and a revived Korean culture after Japanese colonialism, leading to negative perceptions of Japan. Indonesia's "new nation" creation narrative casts colonialism as a negative yet peripheral episode in its national story, resulting in neutral perceptions of the Netherlands.

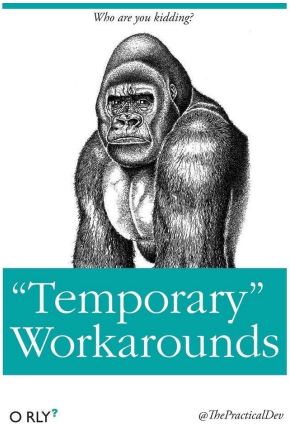
Keywords

Nationalism, Colonialism, Asia, Nation-Building

ODM usage, programming be like...







"*Keep clicking until it works*!"

WHAT IS SENTIMENT ANALYSIS?

Why Sentiment Analysis?

- Quantifies the **evaluative tone** of text — positive, negative, or neutral.
- Turns qualitative impressions into measurable data.
- Enables comparisons **over time, across speakers, or across topics**.
- Especially useful for large text corpora (e.g., textbooks, interviews).

An Example from Korean Textbooks

Evaluative sentences mentioning 일본 (Japan)

- **일본 제국은 조선을 침략하였다.**
  → negative tone (words like 침략, “invasion”)
- **한국은 일본과 협력을 모색하였다.**
  → mildly positive tone (협력, “cooperation”)
- **조선민족은 일본인과의 교류를 경험하였다.**
  → neutral / descriptive tone
- **대한제국은 일본의 근대화를 본받았다.**
  → mixed or ambiguous tone

These illustrate the challenge: subtle variation in tone, even within factual narration.

Basic Pipeline (in Orange Data Mining)

1. **Preprocess Text** — cleaning, tokenization, stopword removal.
2. **Sentiment Analysis** — lexicon-based lookup using multilingual dictionary.
   - Uses Multilingual Sentiment Lexicon.
   - Returns a numeric feature `sentiment` (raw sum of polarity weights).
   - No Bag-of-Words or TF–IDF transformation required.
3. **Filter Zeros** — exclude sentences with no sentiment-bearing words.
4. **Normalize & Visualize** — rescale to comparable range.

Sentences vs. Documents

- **Unit of analysis matters.**
- Tweets or short posts: a single opinion $\Rightarrow$ document-level analysis works.
- Textbooks: thousands of mostly neutral sentences $\Rightarrow$ document-level analysis **dilutes** tone.
- We therefore split full texts into **sentences as mini-documents**.
- Each sentence expresses one idea $\Rightarrow$ clearer evaluative granularity.

Preprocessing for Sentiment

- **Tokenization:** split into morphemes using *KiwiPiePy*.
- **POS filtering:** keep nouns, verbs, adjectives, adverbs.
- **Cleaning:** remove punctuation, numbers, particles, URLs.
- **Goal:** retain only content words carrying evaluation (침략, 자유, 협력).
- **Output:** clean, tokenized column ready for Sentiment Analysis.

Lexicon-Based Sentiment (in ODM)

- Each token matched against dictionary with **polarity weights** (e.g., –5 to +5).
- Sentence score = sum of matched weights.
- 0 = no lexical sentiment detected.
- Transparent, interpretable, multilingual.
- Limitations:
  - Ignores context (negation, irony, historical meaning shifts).
  - Uneven lexicon coverage for Korean and older vocabulary.

Removing Zeros Before Normalization

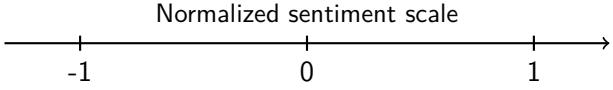
- Zero values = sentences with **no matched sentiment words**.
- Keeping them skews scaling toward 0.
- Use **Select Rows** ⇒ condition: `sentiment` ≠ `0`.
- Then apply *normalization* only to evaluative sentences.

Why Normalize Sentiment?

- Raw scores (–50 to +40) are **sums** of polarity hits.
- Longer or more expressive sentences yield larger magnitudes.
- **Normalization ensures comparability across sentences and periods.**

Some options in Orange

- **Min–max scaling [–1, 1]** (our choice)
  - Easy to interpret; preserves sign.
  - Sensitive to extreme values.
- **Standardization (z-score)**
  - Centers mean at 0; variance = 1.
  - Less sensitive to outliers.



When / When Not to Normalize

Normalize when:

- Comparing sentences across periods or sources.
- Sentence lengths vary widely.
- Preparing data for visualization or teaching.

Skip normalization when:

- Measuring total evaluative load in uniform texts.
- You want to preserve additive meaning (number of sentiment words).

Before and After Normalization

Sentence	Raw Sentiment	Normalized [–1, 1]
일본 제국은 조선을 침략하였다.	–45	–1.00
한국은 일본과 협력하였다.	+12	+0.48
조선민족은 일본인과의 교류를 경험하였다.	0 (removed)	–
대한제국은 일본의 근대화를 본받았다.	+7	+0.30

Common Pitfalls

- Forgetting to filter zeros before normalization $\Rightarrow$ compressed distributions.
- Over-cleaning removes evaluative vocabulary.
- Historical lexicon coverage limits detection.
- Neutral (0) $\neq$ neutral opinion — may mean no lexical signal.
- Always interpret results in context (genre, period, authorship).

VISUALIZING (SENTIMENT) DATA

Why Visualize Sentiment?

- Numbers alone tell us little — visualization helps us **see patterns**.
- We can compare tone **across time periods**, **sources**, or **groups**.
- Key goals:
  - Detect changes in tone over time.
  - Spot outliers (unusually positive or negative sentences).
  - Compare central tendencies (typical sentiment values).

Visualization = turning computation into interpretation.

Box (and Whiskers) Plots

- Summarize how values are distributed.
- The **box** shows the middle 50% of data (interquartile range).
- The **line** inside = median (the typical value).
- The **whiskers** show overall range (without extremes).
- Dots = outliers (unusually high or low scores).

Interpretation:

Taller box = more variation.

Box higher/lower = more positive/negative sentiment.

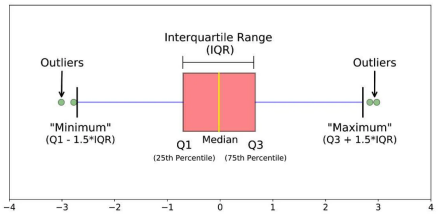
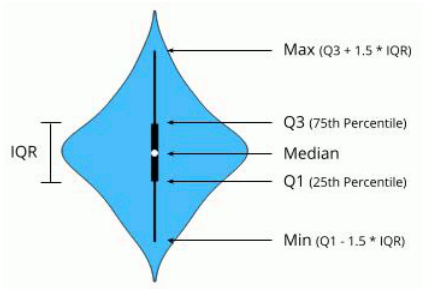


Illustration of box plot structure.

Violin Plots

- Show the **shape** of the distribution.
- Width = how many sentences fall at that sentiment level.
- The wider the violin, the more frequent that value.
- Often include a small box or line for median.
- Reveal patterns (e.g., skew toward positive or negative).



Example of violin plot.

Interpretation:

Symmetric violin = balanced tone.

Wider on one side = skewed sentiment.

ASSIGNMENT