Sentiment Analysis

Applied Text Mining

Dr. Maryam Movahedifar

14-17 July 2025

University of Bremen, Germany
movahedm@uni-bremen.de

Universität Bremen

DATA SCIENCE CENTER

Outline

Introduction to Sentiment Analysis

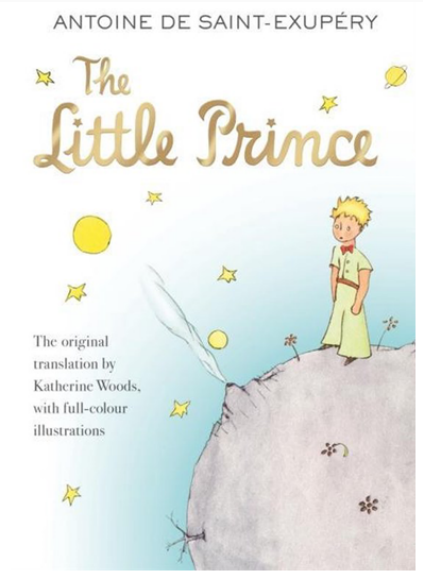
The Little Prince Example

Review Text:

This is a nice book for both young and old. It gives beautiful life lessons in a fun way. Definitely worth the money!

Identified Aspects:

- + Educational
- + Fun
- + Price
- + Funny
- - Readability

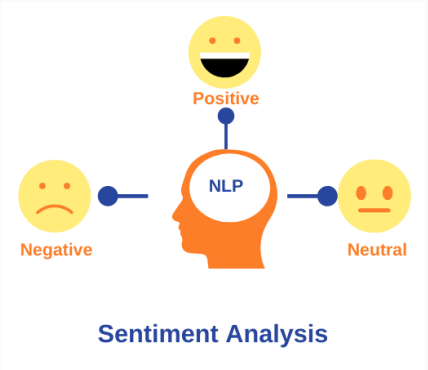


Understanding Sentiment

- **Sentiment** reflects our *feelings*, *attitudes*, and *emotions* toward a subject.
- It is a personal point of view, often driven more by *emotion* than by logical reasoning.
- Importantly, sentiment represents *subjective impressions*, rather than objective facts.

What is Sentiment Analysis?

- Sentiment Analysis uses **(NLP)** to automatically identify and classify emotions expressed in text.
- It transforms *unstructured* data—like social media posts, reviews, or blogs—into actionable insights.
- Known by various names: *Opinion Mining*, *Sentiment Mining*, and *Sentiment Classification*.



Sentiment Analysis Challenges

Sarcasm and irony

Contextual understanding

Language variations

Data quality

Privacy concerns

Real-World Applications of Sentiment Analysis

- **Book Reviews:** Classify opinions as positive or negative.
- **Cultural Studies:** Analyze sentiment in historic plays or literature.
- **Product Feedback:** Gauge public opinion on new launches, e.g., the latest smartphone.
- **Social Issues:** Understand attitudes on immigration, politics, and more.
- **Entertainment:** Evaluate movie reviews on platforms like Netflix or IMDB.
- **Marketing:** Measure consumer confidence and brand sentiment.
- **Healthcare:** Assess patient satisfaction from hospital feedback.
- **Social Media:** Track trending moods and topics in real time.

Opinion Types and Challenges

Opinion Types

Regular Opinions

These are opinions about a specific entity or aspect.

Direct Opinion: "The touch screen is really cool."

Indirect Opinion: "After taking the drug, my pain has gone."

Comparative Opinions

These involve comparing two or more entities.

Example: "iPhone is better than Blackberry."

Sentiment Analysis Task Summary

Basic Task:

Decide whether the sentiment in a text is *positive* or *negative*.

Intermediate Tasks:

- Include a third option: *neutral*.
- Use numerical scales (e.g., 1 to 5) to capture strength of sentiment.

Advanced Tasks:

- Identify the **target** — what the sentiment refers to.
- Identify the **source** — who is expressing the opinion.
- Detect **comparisons** or complex sentiment.
- Understand **implicit sentiment** that isn't directly stated.

NLP Challenges in Sentiment Analysis

Limitations of Bag of Words:

Ignores word order, syntax, and nuance in text.

Subtle Sentiment Expression:

Irony: "What a great car, it stopped working on the second day."

Neutral language: "The concert didn't meet my expectations."

Context and Domain Dependence:

Same phrase can have different sentiment:

"Long queue" (negative) vs. "Long battery life" (positive)

Syntax and Negation:

Word order and negations can completely change meaning.

Methods for Sentiment Analysis

Methods for Sentiment Analysis

Lexicon-Based Methods

- **Dictionary-Based:**
Uses predefined sentiment word lists like *good*, *awesome*, or *terrible*.
- **Corpus-Based:**
Builds or expands lexicons by analyzing how words co-occur in large text collections.

Supervised Learning Methods

- **Traditional Machine Learning:**
Uses models like *Naïve Bayes* or *SVM*, trained on labeled datasets.
- **Deep Learning:**
Employs powerful models like *BERT*, *GPT*, etc., which capture deeper context and subtle sentiment.

Lexicon-based Methods

What Are Sentiment Lexicons?

- Lexicons are lists of words paired with **sentiment scores**.
- Scores come in two types:
  - Binary: Positive (1) or Negative (-1)
  - Intensity: Scores that show how strong the sentiment is, often from 0 to 1 or scaled.
- Lexicons cover various categories:
  - Positive and negative words
  - Emotions and feelings
  - Negation words that can flip meaning

```
brainwashing    -3
brave    2
breakthrough    3
breathtaking    5
bribe    -3
bright   1
brightest       2
brightness      1
brilliant       4
brisk    2
broke    -1
broken   -1
```

Basic Lexicon Approach

How it works:

- Sentiment is measured on two independent scales:
  - Positive scores: {1, 2, ..., 5}
  - Negative scores: {-5, -4, ..., -1}
- Helps handle sentences that mix positive and negative feelings.

Example: *"He is brilliant but boring"*

- `brilliant` → +4
- `boring` → -2
- Overall sentiment: +2 (leans positive)

VADER Sentiment Analysis

VADER = Valence Aware Dictionary and sEntiment Reasoner

- A lexicon and rule-based sentiment analysis tool designed for social media text.
- Sentiment scores range from **-4 (very negative)** to **+4 (very positive)**.

VADER uses five smart rules (heuristics):

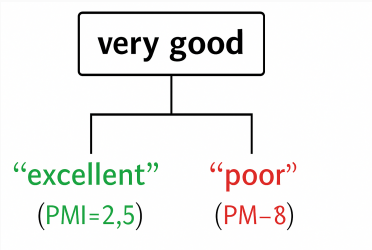
- Punctuation: More exclamation marks = stronger emotion.
- Capitalization: ALL CAPS increase intensity.
- Degree Modifiers: Words like "very" or "extremely" amplify meaning.
- Contrastive Conjunction "But": Flips sentiment focus.
- Negation Handling: Detects nearby negating words like "not" or "never".

Measuring the Polarity of a Phrase

Key Insight:

Positive and negative phrases tend to occur near different reference words.

- Positive phrases often appear near "excellent".
- Negative phrases often appear near "poor".
- We use Pointwise Mutual Information (PMI) to measure these relationships statistically.

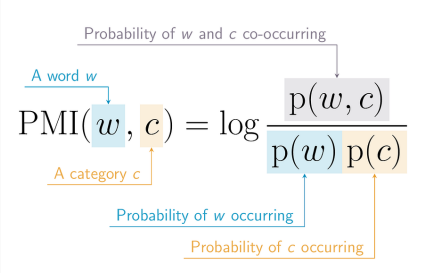


Pointwise Mutual Information (PMI)

What does PMI measure?

PMI tells us how strongly two words are associated — for example, whether a phrase appears more often near positive words like excellent or negative ones like poor.

- High PMI $\rightarrow$ words co-occur more than expected.
- Positive PMI suggests stronger association with the reference word.
- Used to detect sentiment polarity in context.
- Based on word frequencies from large corpora.



Supervised Methods

Supervised Methods for Sentiment Analysis

- **Key Steps:**
  - Pre-processing & tokenization
  - Feature representation & selection
  - Classification
  - Evaluation
- **Tokenization Challenges:**
  - Handling HTML/XML and Twitter syntax (hashtags, @mentions)
  - Preserving capitalization (ALL CAPS matters!)
  - Managing emoticons, phone numbers, dates
- **Tools:** Potts sentiment tokenizer, O'Connor Twitter tokenizer
- **Watch Out:** Stemming pitfalls — e.g., "**objective**" vs "**objection**"

Features and Negation Handling

- **Key features:**
  - Term frequencies, POS tags, opinion words, negations
  - Stylistic and syntactic dependency features

- **Negation handling:**
  - Add "NOT_" prefix to words between negation and punctuation
  - Example: "wasn't terrible" flips polarity
  - Negation effect varies in intensity

- **Advanced:** Kiritchenko et al. (2014) — separate lexicons for negated and affirmative words

Pros and Cons of Supervised Sentiment Analysis

Advantages

- High accuracy on many tasks
- Adapts well to domain-specific data
- Results can be interpretable

Disadvantages

- Needs a lot of labeled training data
- Context handling is still limited
- Feature engineering can be time-consuming
- Struggles with multiclass sentiment tasks

Practical