Sentiment Analysis Explained
What is Sentiment Analysis?

Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. In addition, data analytics companies often integrate third-party sentiment analysis APIs into their own customer experience management, social media monitoring, or workforce analytics products, in order to deliver useful insights to their own customers.

This white paper will explain how basic sentiment analysis works, evaluate the advantages and drawbacks of rules-based sentiment analysis, and outline the role of machine learning in sentiment analysis. Finally, we’ll explore the top applications of sentiment analysis before concluding with some helpful resources for further learning.
THE BASICS

How does sentiment analysis work?
Basic sentiment analysis of text documents follows a straightforward process:

1. Break each text document down into its component parts
2. Identify each sentiment-bearing phrase and component
3. Assign a sentiment score to each phrase and component (-1 to +1)
4. Optional: Combine scores for multi-layered sentiment analysis

As you’ll see, the underlying technology is very complicated. But for a simple explanation, consider these sentences:

Terrible pitching and awful hitting led to another crushing loss.
Bad pitching and mediocre hitting cost us another close game.

Both sentences discuss a similar subject, the loss of a baseball game. But you, the human reading them, can clearly see that the first sentence’s tone is much more negative.

Your brain figures this out by looking for and interpreting sentiment-bearing phrases — that is, words and phrases that carry a tone or opinion. These usually appear as adjective-noun combinations. In the examples above, the sentiment-bearing phrases are:

Terrible pitching  I  awful hitting  I  crushing loss
Bad pitching  I  mediocre hitting  I  close game
You have encountered words like these many thousands of times over your lifetime across a range of contexts. And from these experiences, you've learned to understand the relative strength of each adjective, receiving input and feedback along the way from teachers and peers.

When you read the sentences like the examples before, your brain draws on your accumulated knowledge to identify each sentiment-bearing phrase and interpret their negativity or positivity. Usually this happens subconsciously. For example, you instinctively know that a game that ends in a “crushing loss” has a higher score differential than the “close game,” because you understand that “crushing” is a stronger adjective than “close.”

So, why are we using baseball games to explain how human brains do sentiment analysis? The answer is simple: **computer sentiment analysis works (almost) the same way.**
What is a sentiment library?

Much in the way your brain remembers the descriptive words you encounter over your lifetime and their relative “sentiment weight,” a basic sentiment analysis system draws on a sentiment library to understand the sentiment-bearing phrases it encounters. Sentiment libraries are very large collections of adjectives (good, wonderful, awful, horrible) and phrases (good game, wonderful story, awful performance, horrible show) that have been hand-scored by human coders.

This manual sentiment scoring is a tricky process, because everyone involved needs to reach some agreement on how strong or weak each score should be relative to the other scores. If one person gives “bad” a sentiment score of -0.5, but another person gives “awful” the same score, your sentiment analysis system will conclude that both words are equally negative.

What’s more, a multilingual sentiment analysis engine must maintain unique libraries for each language it supports. And each of these libraries must be maintained constantly: scores tweaked, new phrases added, irrelevant phrases removed.

How does a simple rules-based sentiment analysis system work?

Once the sentiment libraries are prepared, software engineers write a series of guidelines (“rules”) to help the computer evaluate the sentiment expressed towards a particular entity (noun or pronoun) based on its nearness to known positive and negative words (adjectives and adverbs). To continue our baseball example, an engineer might create search rules that look like:

\[(\text{pitching}) \text{ NEAR } (\text{good, wonderful, spectacular})\]
\[(\text{pitching}) \text{ NEAR } (\text{bad, horrible, awful})\]

These queries return a “hit count” representing how many times the word “pitching” appears near each adjective. The system then combines these hit counts using a complex mathematical operation called a “log odds ratio.” The outcome is a numerical sentiment score for each phrase, usually on a scale of –1 (very negative) to +1 (very positive).

This is a simplified example, but it serves to illustrate the basic concepts behind rules-based sentiment analysis.
What is the role of Part of Speech tagging in sentiment analysis?

Before you can analyze a sentence and phrase for sentiment, however, you need to understand the pieces that form it. The process of breaking a document down into its component parts involves several sub-functions, including Part of Speech (PoS) tagging. Part of Speech tagging is the process of identifying the structural elements of a text document, such as verbs, nouns, adjectives, and adverbs.

Most languages follow some basic rules and patterns that can be coded into a computer program to build a basic Part of Speech tagger. In English, for example, a number followed by a proper noun and the word “Street” most often denotes a street address. A series of characters interrupted by an @ sign and ending with “.com,” “.net,” or “.org” usually represents an email address. People’s names often follow generalized two- or three-word patterns of nouns.

Nouns and pronouns are most likely to represent named entities, while adjectives and adverbs usually describe those entities in emotion-laden terms. By identifying adjective-noun combinations, such as “terrible pitching” and “mediocre hitting,” a sentiment analysis system gains its first clue that it’s looking at a sentiment-bearing phrase. Of course, not every sentiment-bearing phrase takes an adjective-noun form. “Cost us,” from the example sentences earlier, is a noun-pronoun combination but bears some negative sentiment.

Accurate part of speech tagging is critical for reliable sentiment analysis, so it’s important that a rules-based system account for these variations.
DIVING DEEPER

How negators and intensifiers affect sentiment analysis

Consider a hotel review that reads:

The bed was super comfy. The chair wasn’t bad, either.

A simple rules-based sentiment analysis system will see that “comfy” describes “bed” and give the entity in question a positive sentiment score. But that score will be artificially low, even if it’s technically correct, because the system hasn’t considered the intensifying adverb “super.” When a customer likes their bed so much, the corresponding entity’s sentiment score should reflect that intensity.

Even worse, the same system is likely to think that “bad” describes “chair.” This overlooks the key word “wasn’t,” which negates the negative implication and should change the sentiment score for “chairs” to positive or neutral.

Here’s another example:

The food was pretty good, but I won’t bother coming back.

A simple rules-based sentiment analysis system will see that “good” describes “food,” slap on a positive sentiment score, and move on to the next review.

But you (the human reader) can see that this review actually tells a more nuanced story. Even though the writer liked their food, something about their experience turned them off. This review illustrates why an automated sentiment analysis system must consider negators and intensifiers as it assigns sentiment scores.
Drawbacks of rules-based sentiment analysis

The simplicity of rules-based sentiment analysis makes it a good option for basic document-level sentiment scoring of predictable text documents, such as limited-scope survey responses. However, a purely rules-based sentiment analysis system has many drawbacks that negate most of these advantages.

A rules-based system must contain a rule for every word combination in its sentiment library. Creating and maintaining these rules requires tedious manual labor. And in the end, rules can't hope to keep up with the evolution of natural human language. Instant messaging has butchered the traditional rules of grammar, and no ruleset can account for every abbreviation, acronym, double-meaning and misspelling that may appear in any given text document.

In addition, a rules-based system that fails to consider negators and intensifiers is inherently naïve, as we've seen. Out of context, a document-level sentiment score can lead you to draw false conclusions. Lastly, a purely rules-based sentiment analysis system is very delicate. When something new pops up in a text document that the rules don't account for, the system can't assign a score. In some cases, the entire program will break down and require an engineer to painstakingly find and fix the problem with a new rule.

In the end, anyone who requires nuanced analytics, or who can't deal with ruleset maintenance, should look for a tool that also leverages machine learning.
Multi-layered sentiment analysis and why it is important

Imagine a Yelp review that says:

The linguini was great, but the room was way too dark.

In this document, “linguini” is described by “great,” which deserves a positive sentiment score. But “room” is described negatively. Depending on the exact sentiment score each phrase is given, the two may cancel each other out and return neutral sentiment for the larger document.

As this example demonstrates, document-level sentiment scoring paints a broad picture that can obscure important details. In this case, the restaurant manager misses the crucial insight that she may be losing repeat business because customers don’t like her dining room’s ambience.

Multi-layered sentiment analysis systems solve this problem by assigning sentiment scores not just to documents, but to individual entities, topics, themes and categories as well. The result is a system that, given the same Yelp review, can return:

| Entity sentiment: “linguini” ( +0.5 ) |
| Entity sentiment: “room” ( -0.25 ) |
| Theme sentiment: “dark” ( -0.25 ) |
| Category sentiment: “dining” ( +0.25 ) |

In this case, the positive entity sentiment of “linguini” and the negative sentiment of “room” would partially cancel each other out to influence a neutral sentiment of category “dining.”

This multi-layered analytics approach reveals deeper insights into the sentiment directed at individual people, places, and things, and the context behind those opinions.
**M A C H I N E L E A R N I N G**

**How is machine learning used for sentiment analysis?**

The primary role of machine learning in sentiment analysis is to **improve and automate** the low-level text analytics functions that sentiment analysis relies on. For example, data scientists can train a machine learning model to identify nouns by feeding it a large volume of text documents containing pre-tagged examples. Using supervised and unsupervised machine learning techniques, such as neural networks and deep learning, the model will learn what nouns “look like.”

Once the model is ready, the same data scientists can apply those training methods towards building new models to identify other parts of speech. The result is quick and reliable Part of Speech tagging that helps the larger text analytics system identify sentiment-bearing phrases more effectively.

Machine learning also helps data analysts solve tricky problems caused by the inconsistencies of natural language. For example, the word “sick” can mean very different things depending on the context it’s used in. Creating a sentiment analysis ruleset to account for every possible meaning is impossible.

But if you feed a machine learning model with a few thousand pre-tagged examples, the system can learn what “sick burn” means in the context of video gaming, versus in the context of healthcare. And you can apply similar training methods to understand other double-meanings as well.

**What is a hybrid sentiment analysis system?**

Hybrid sentiment analysis systems combine machine learning with traditional rules to make up for the deficiencies of each approach. Rules-based sentiment analysis, for example, can be an effective way to build a foundation for PoS tagging and sentiment analysis. But as we’ve seen, these rulesets quickly grow to become unmanageable. This is where machine learning can step in to shoulder the load of complex natural language processing tasks, such as understanding double-meanings.

Most hybrid sentiment analysis systems combine machine learning with software rules across the entire text analytics function stack, from low-level tokenization and syntax analysis all the way up to the highest-levels of sentiment analysis.
APPLICATIONS

How is sentiment analysis used?

Broadly speaking, sentiment analysis is most effective when used as a tool for Voice of Customer and Voice of Employee. Business analysts, product managers, customer support directors, human resources and workforce analysts, and other stakeholders use sentiment analysis to understand how customers and employees feel about particular subjects, and why they feel that way.

Sentiment analysis for Voice of Customer

In the age of social media, a single viral review can burn down an entire brand. On the other hand, research by Bain & Co. shows that good experiences can grow 4-8% revenue over competition by increasing customer lifecycle 6-14x and improving retention up to 55%.

Automated sentiment analysis tools are the key drivers of this growth. By analyzing tweets, online reviews and news articles at scale, business analysts gain useful insights into how consumers feel about their brands, products and services. Customer support directors and social media managers flag and address trending issues before they go viral, while forwarding these pain points to product managers to make informed feature decisions.

Sentiment analysis for Voice of Employee

The cost of replacing a single employee averages 20-30% of salary, according to the Center for American Progress. Yet 20% of workers voluntarily leave their jobs each year, while another 17% are fired or let go. To combat this issue, human resources teams are turning to data analytics to help them reduce turnover and improve performance.

Sentiment analysis helps workforce analysts and HR directors cut off employee churn at the source by understanding what employees are discussing and how they feel. By analytics of employee surveys, Slack messages, emails, and other communications, HR teams get the info they need to proactively address pain points and improve morale.

1 Bain and Co., Insights Infographic; http://www2.bain.com/infographics/five-disciplines/
FURTHER READING

Where can I learn more about sentiment analysis?

• Explore the difference between Machine Learning and Natural Language Processing: https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1

• Take a Coursera online course on Text Mining and Analytics: https://www.coursera.org/learn/text-mining


Where can I try sentiment analysis for free?

• Take a guided, interactive demo of multi-layered sentiment analysis dashboards: https://www.lexalytics.com/ssv

• Get started with the Stanford NLP or Natural Language Toolkit (NLTK) open source distributions: https://nlp.stanford.edu/software/ http://www.nltk.org/

• Play with a web demo to see some basic sentiment analysis in action: https://www.lexalytics.com/demo
Lexalytics processes BILLIONS of unstructured documents every day, GLOBALLY.

We transform unstructured text into usable data and powerful stories.

Our on-premise Salience® engine, SaaS Semantria® API, and end-to-end Lexalytics Intelligence Platform® combine natural language processing with artificial intelligence to reveal context-rich patterns and insights within comments, reviews, surveys, and other text documents.

Data analytics and data analyst companies rely on Lexalytics to build better products, share insights between engineering, marketing, PR, and support teams, and drive business growth.

For more information, visit www.lexalytics.com or call 1-800-377-8036