NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a collection of techniques that allows computers to work with and analyze text data. To describe a text, it is first broken into tokens, which can be words or subwords (smaller meaningful units of words). These tokens can then be summarized using bar charts of the most frequent tokens or visualized using a word cloud, where more frequently used tokens appear larger.

For supervised text analysis, relatively small samples of text (up to 512 tokens) are paired with a target variable. For example, excerpts from different authors may be labeled by author, product reviews may be labeled from 1 to 5 stars, or texts may be labeled by sentiment (negative, neutral, or positive). Common sentiment analysis examples include news headlines or Twitter reviews.

Classical statistical methods such as cumulative logistic regression can be used to model ordinal sentiment outcomes (e.g., 1–5 star ratings) by regressing the sentiment score on features such as the 100 most frequently used words. Words with negative regression coefficients are associated with more negative sentiment, whereas words with positive coefficients are associated with more positive sentiment.

A modern machine learning approach for text analysis is the Bidirectional Encoder Representations from Transformers (BERT) model. It processes all words in a sentence simultaneously (in both directions) and learns how strongly each word relates to every other word in the sentence. This attention mechanism allows the model to capture long-range dependencies and contextual relationships efficiently.

The key theoretical idea behind BERT is bidirectional contextual learning. During pretraining, BERT uses a task called masked language modeling, where some words in a sentence are randomly hidden and the model learns to predict them using both left and right context. After pretraining on large amount of text, BERT can be used for specific tasks such as classification of new texts.