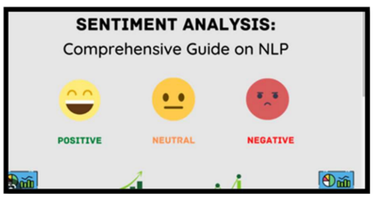
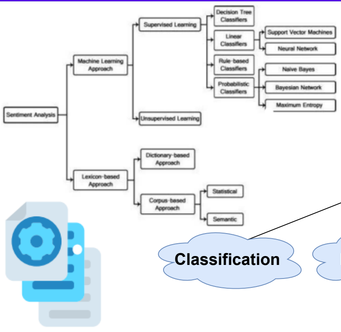
SENTIMENTAL ANALYSIS MOVIE REVIEWS



The problem statement of the article is to predict the polarity (positive or negative) of movie reviews by understanding the meaning and relationship between words.The goal is to calculate the sentiment of sentences extracted from movie reviews and determine if they are positive, negative, or neutral.

Data Collection & Material

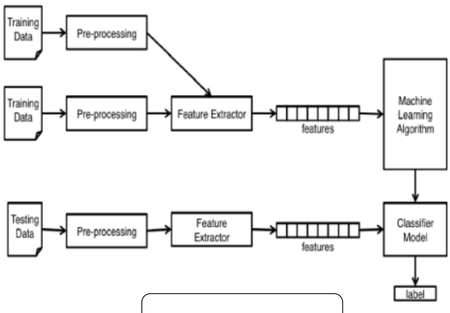


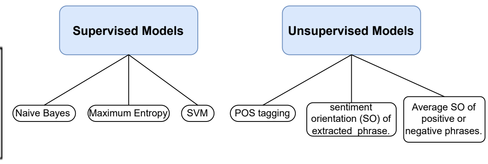
IMDB Movie Review dataset
The variables captured in the dataset are likely the text of the movie reviews and their corresponding sentiment labels (positive or negative).

Proposed Approach

- Classification
- Models
- Feature Extraction
- Sentiment Technique

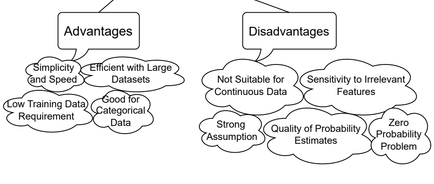
Models Implemented

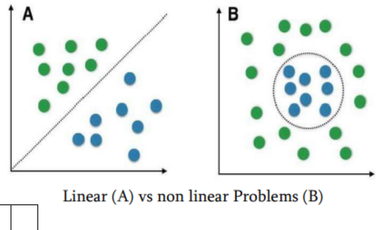




- Naive Bayes assumes features in a dataset which are mutually independent.
- Naive Bayes can perform well, especially with small sample sizes such as in case of spam filtering.
- Strong violations of the independence assumption or non-linear classification problems can lead to poor performance with Naive Bayes.
- Compare models for accuracy and efficiency.
- Linear problems can be separated by straight lines or hyperplanes, making Naive Bayes suitable.
- Non-linear problems require non-linear classifiers like instance-based nearest neighbor classifiers.
- The probability model of Naive Bayes classification and applies it to classify messages (spam or not) using a publicly available SMS dataset in Python.

Naive Bayes Classification





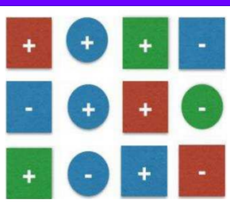
Naive Bayes Classifier (Conditional probabilities)

Independence and Identically Distributed (i.i.d.) Assumption:
In the context of Bayes classifiers, it is assumed that the samples are independent and identically distributed (i.i.d.).
Independence means that the probability of one observation (sample) does not affect the probability of another observation. This assumption is fundamental for the Bayes classifier.

Conditional Independence of Features
The class-conditional probabilities or likelihoods of the samples can be directly estimated from the training data.
For a d-dimensional feature vector x, the class conditional probability P(x|ωj) can be calculated as the product of individual feature probabilities: P(x1|ωj) * P(x2|ωj) * ... * P(xd|ωj), where ωj is the class.

Maximum-Likelihood Estimate for Categorical Data
When dealing with categorical data (such as text classification or categorical features), the individual likelihoods for each feature vector feature can be estimated using the maximum-likelihood estimate.
The maximum-likelihood estimate for P(xi|ωj) is calculated as: P^(xi|ωj) = N(xi, ωj) / N(ωj), where N(xi, ωj) is the number of times feature xi appears in samples from class ωj, and N(ωj) is the total count of all features in class ωj.

Observations & Results



It can be inferred that the Naive Bayes classifier was used to predict the polarity of movie reviews as positive or negative. The article suggests that mining movie reviews and generating valuable metadata can provide insights into the general sentiment around movies in an independent way

Conclusion & Recommendation

Sentiment analysis plays a crucial role in enhancing modern businesses by harnessing the power of Natural Language Processing (NLP) to decipher human language and make it accessible to machines. Additionally, Machine Learning (ML) significantly contributes to the effectiveness of sentiment analysis, making it the top choice in the market. The algorithms driving sentiment analysis not only offer a comprehensive understanding of data but also provide valuable solutions to real-world challenges.

In the future, Generative Pre-Trained Transformer (GPT) is the best in case anybody uses NLP in the AI. It helps the language of a pre-trained model that can help various tasks to be fine-tuned and better apply the appropriate model, increasing the accuracy level.