Text Classification

Ch.5 Text Analytics with Python. Dipanjan Sarkar. 2019. Apress
Ch.2 NLP with Python. Steven Bird et al. 2009. O'Reilly Media
Ch.1 Mastering NLP with Python. Deepti Chopra et al. 2016. Packt

Text Classification Definition

D - a set of records $\{X_1,\ldots,X_N\}$

C - a set of labels $\{c_1 \ldots c_n\}$

T – a Text classification system

$$\boxed{\mathbf{T} : \mathbf{D} \rightarrow \mathbf{C}_x}$$

The training model predicts a class label

(Dipanjan Sarkar. 2019. Ch.5)

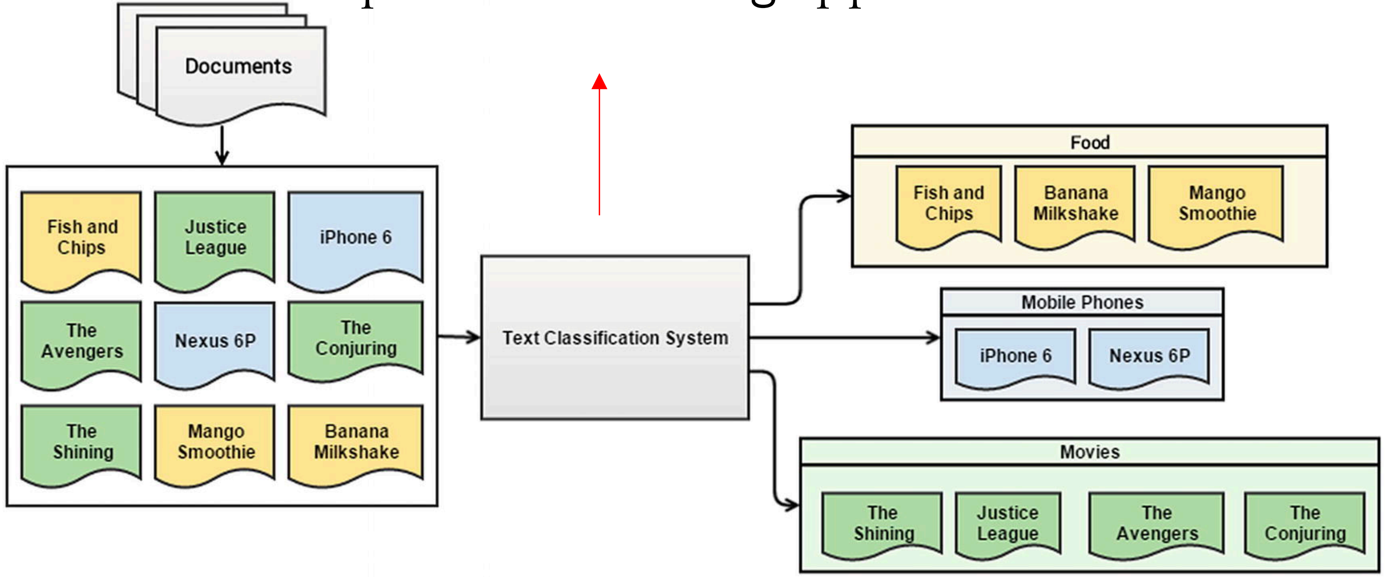
Applications

News Filtering and Organization

Document Organization and Filtering

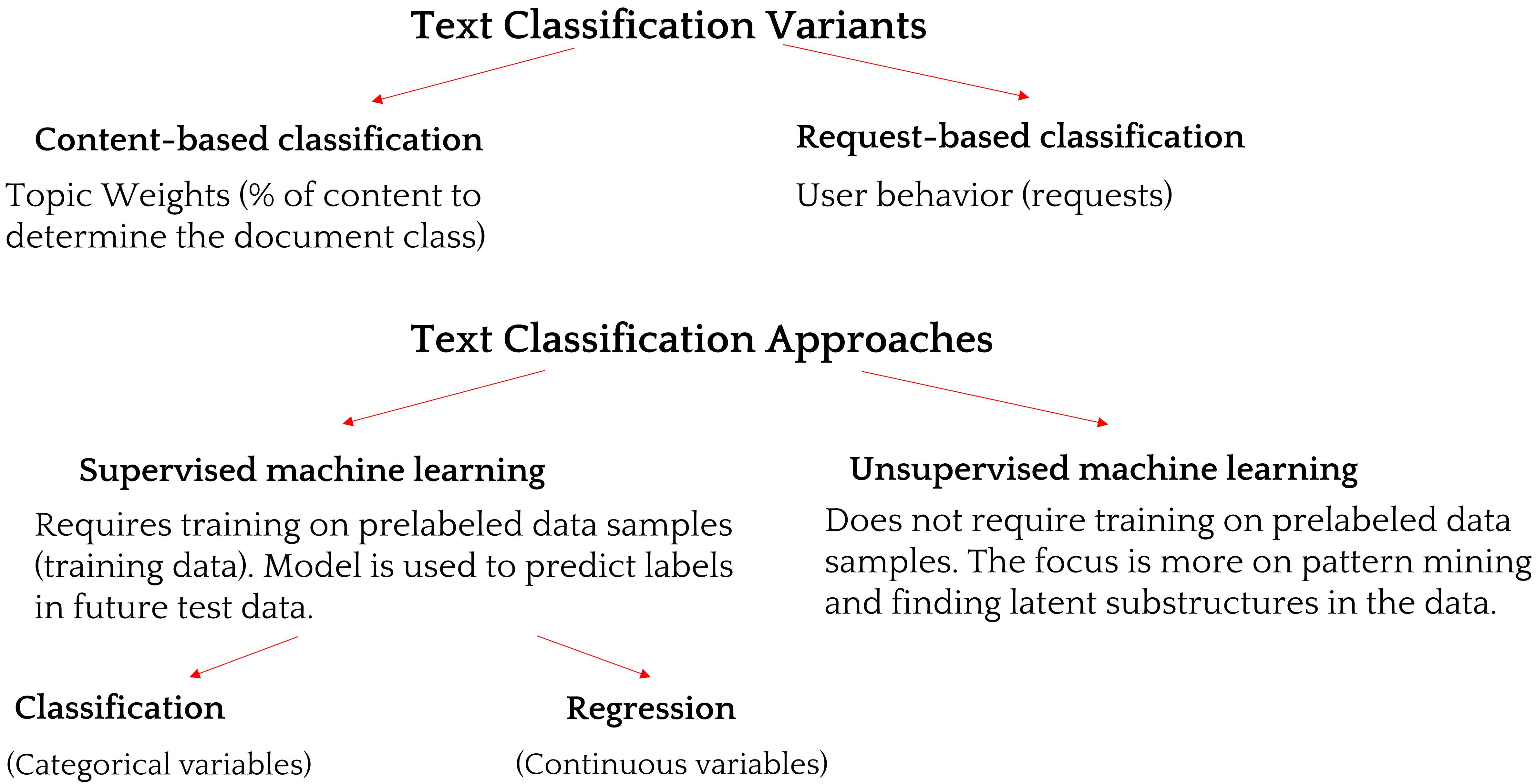
Sentiment analysis

Supervised learning approach



(Charu Aggarwal, 2014, Ch.11)

Text Classification



Text Classification Variants

Content-based classification

Topic Weights (% of content to determine the document class)

Request-based classification

User behavior (requests)

Text Classification Approaches

Supervised machine learning

Requires training on prelabeled data samples (training data). Model is used to predict labels in future test data.

Classification

(Categorical variables)

Regression

(Continuous variables)

Unsupervised machine learning

Does not require training on prelabeled data samples. The focus is more on pattern mining and finding latent substructures in the data.