Session 49

Classification for Text Dataset

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

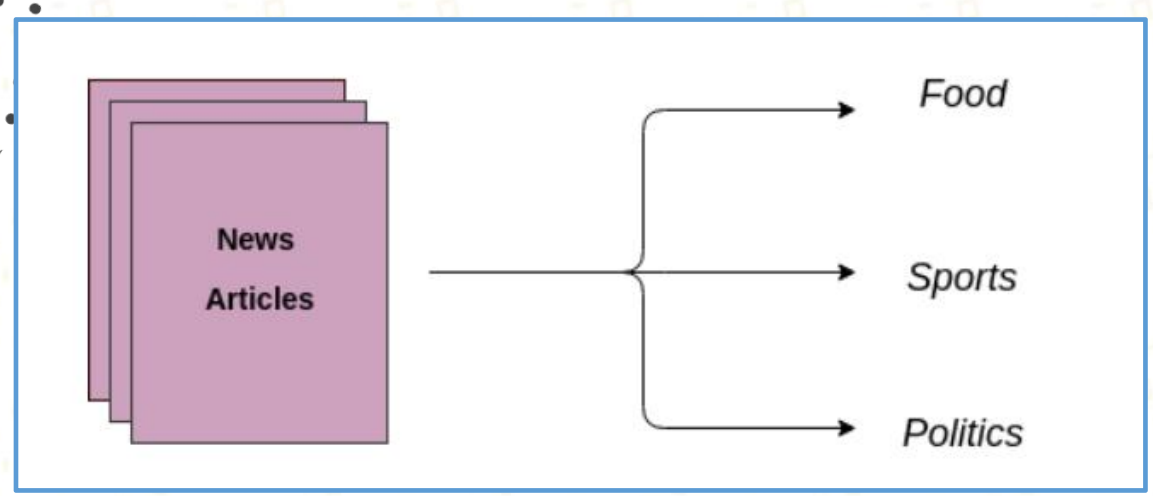
1. Text Classification
2. Sentiment analysis
3. Classification using Machine Learning models

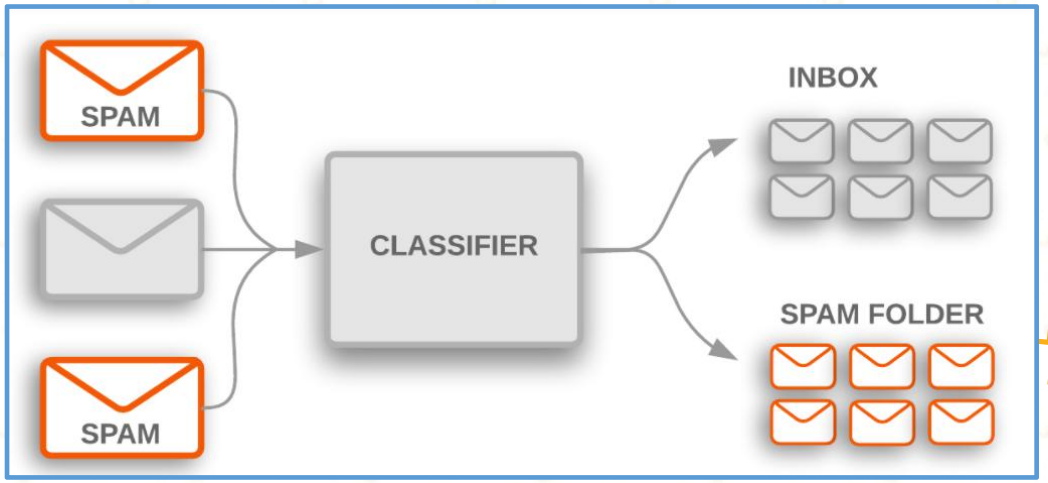
DigitalSkola
Uncover The World Of Digital Skills With Us

Text Classification

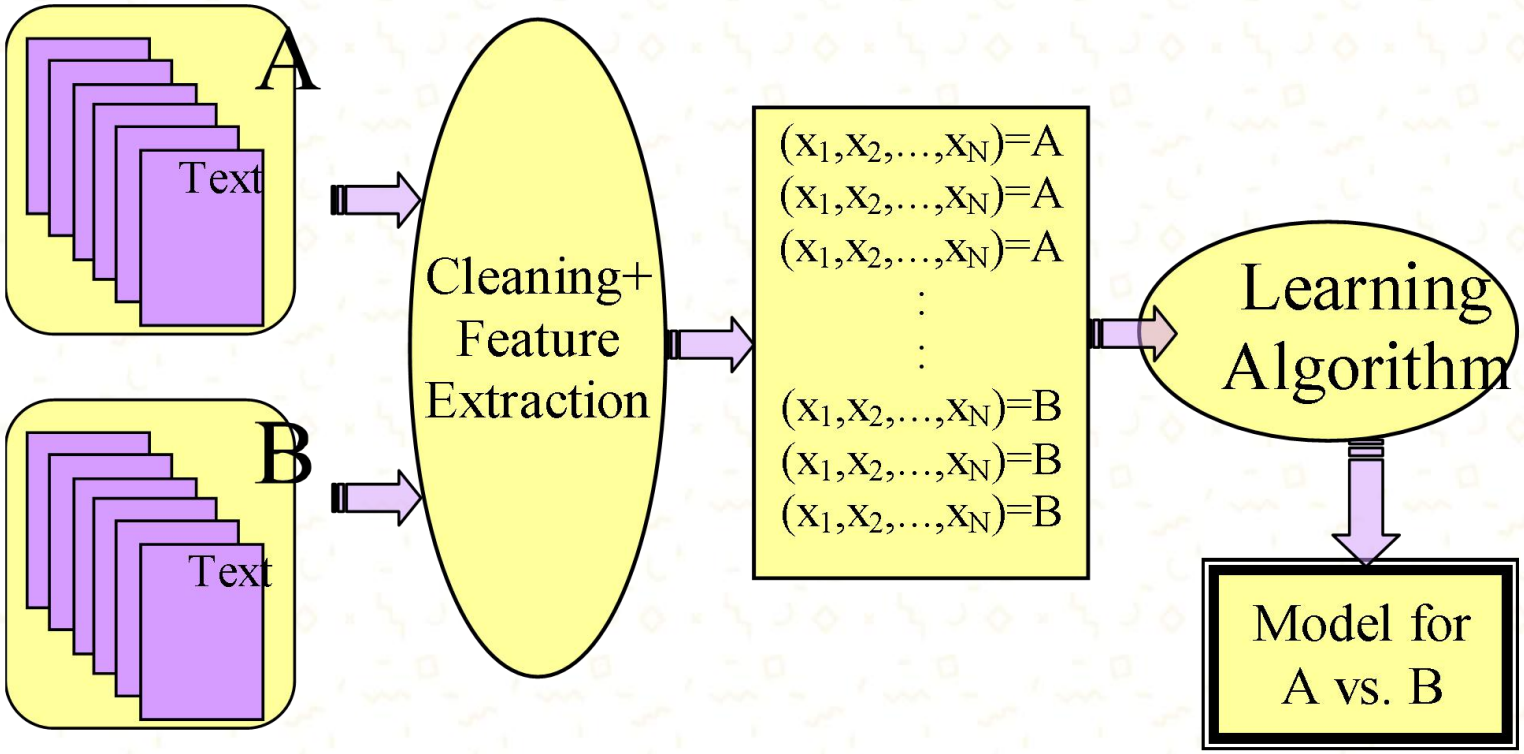
Text Classification

- *Text classification* juga dikenal sebagai *text tagging* atau *text categorization* adalah proses mengkategorikan teks ke dalam kelompok tertentu.
- *Text classification* adalah salah satu tugas dasar dalam *natural language processing (NLP)* dengan aplikasi yang luas contohnya *sentiment analysis, topic labeling, spam detection, dan intent detection*.



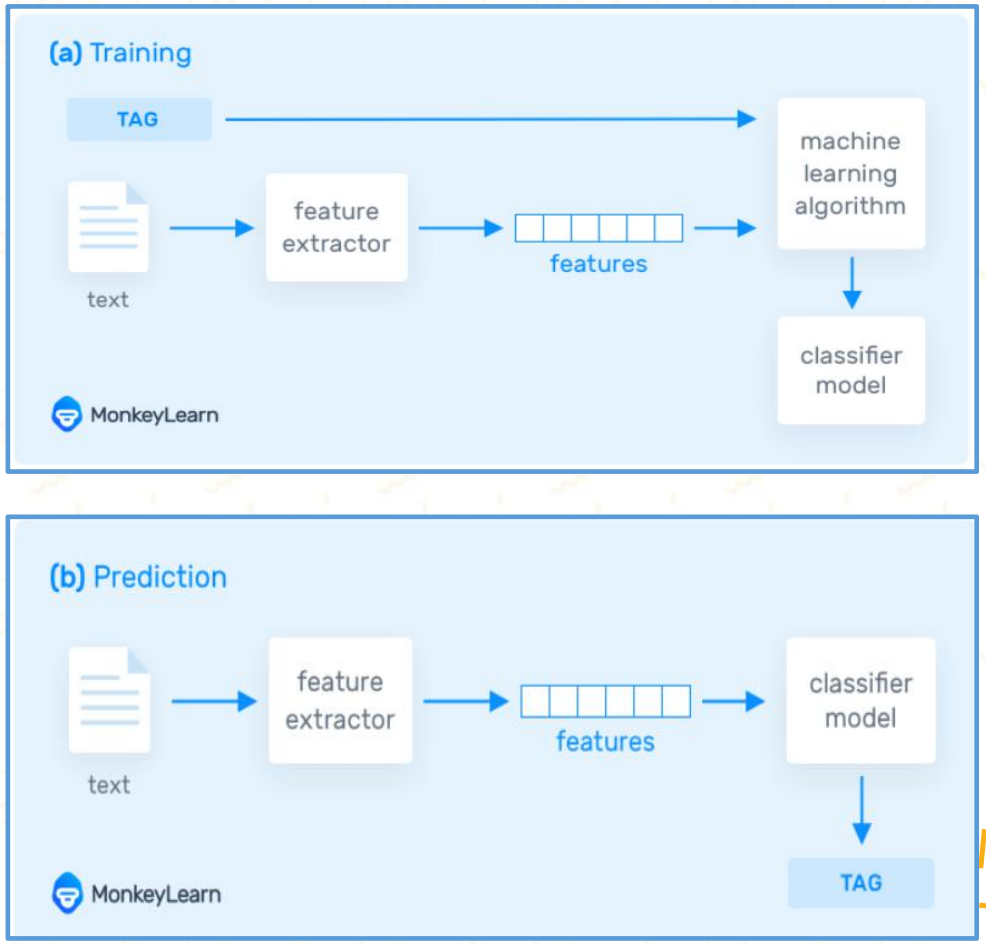


Text Classification



Text Classification

- Ada tiga pendekatan dalam *text classification*
- ***Rule-based System***
  - Teks dipisahkan ke dalam kelompok terorganisir menggunakan *handicraft linguistic rules.*
- ***Machine Learning-based System***
  - *ML-based classifier* membuat klasifikasi berdasarkan pengamatan sebelumnya dari kumpulan data
- ***Hybrid System***
  - Menggabungkan *machine learning classifier* dengan *rule-based system*, digunakan untuk meningkatkan performa.



Machine Learning-based System

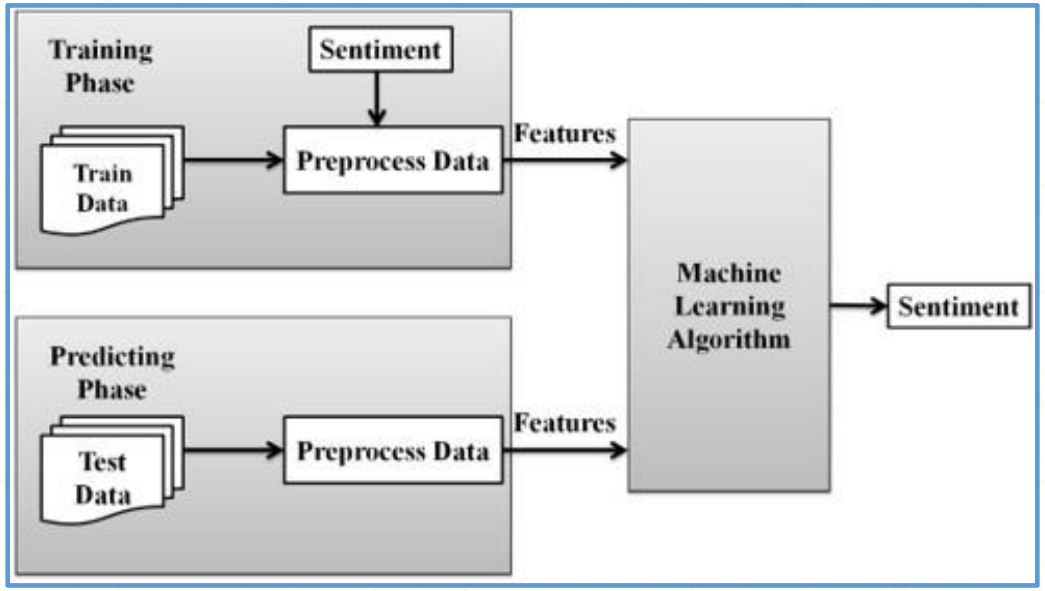
https://monkeylearn.com/text-classification/

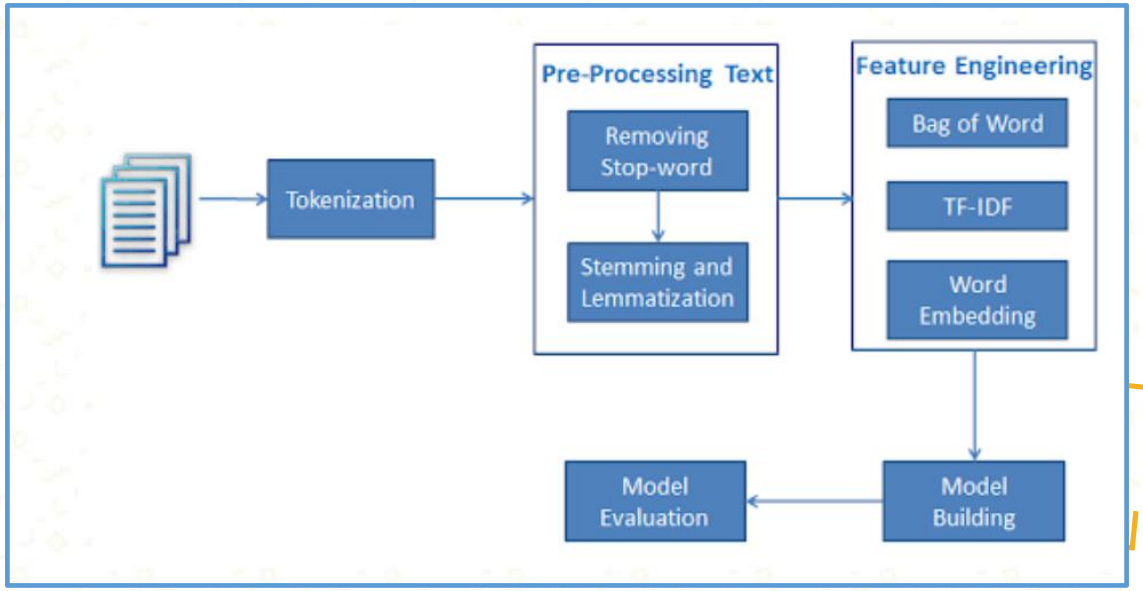
DigitalSkola
Uncover The World Of Digital Skills With Us

Sentiment analysis

Sentiment Analysis-Definition

- Salah satu contoh aplikasi dari *text classification* adalah *sentiment analysis.*
- Adalah metode yang secara otomatis memahami persepsi pelanggan terhadap suatu produk atau layanan berdasarkan komentar mereka.





Pre-processing the Text

1. Tokenization
2. Removing stop words
3. Stemming or Lemmatization

sudah dijelaskan di pertemuan sebelumnya

1

Text
"The cat sat on the mat."
↓
Tokens
"the", "cat", "sat", "on", "the", "mat", "."

2

Stop Words

These words include:

- a
- I
- the
- in
- of
- for
- at
- to
- on
- with
- from

3

Stemming vs Lemmatization

Feature extraction

1. Bag of Words
2. TF-IDF (Term frequency–inverse document frequency)
3. Word Embedding

sudah dijelaskan di pertemuan sebelumnya

```
doc1 = "saya belajar pemrograman dan belajar melukis"
doc2 = "saya membantu adik saya belajar menulis"
doc3 = "ibu belajar menjahit"
```

1

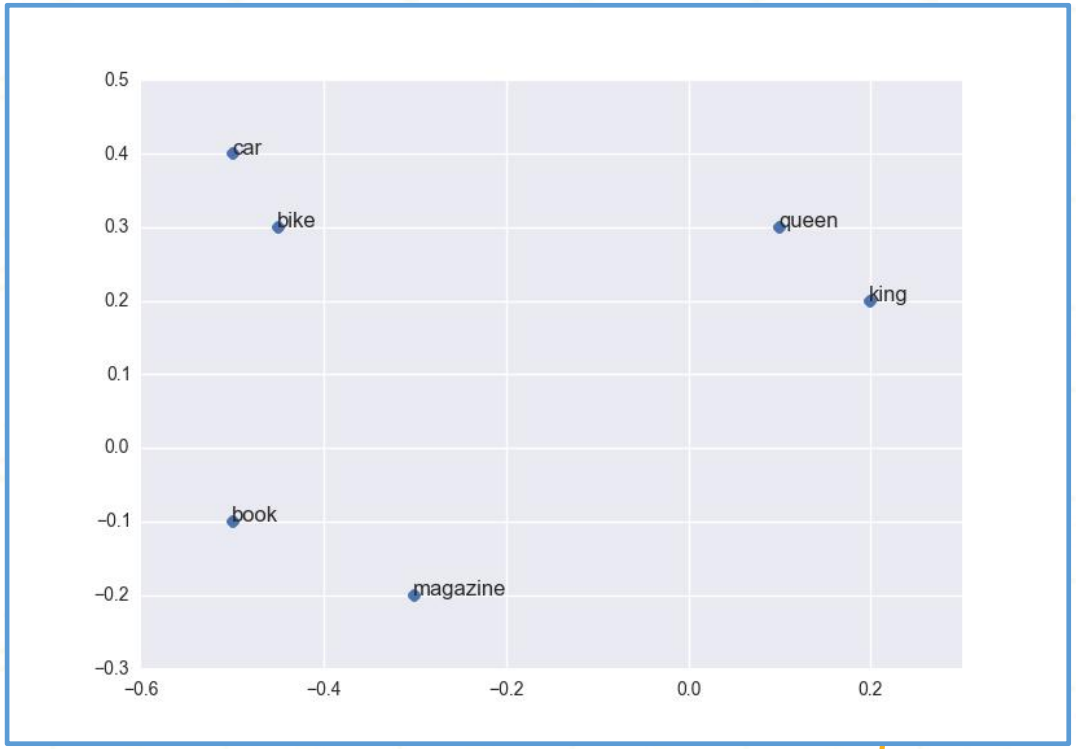
adik	belajar	dan	ibu	melukis	membantu	menjahit	menulis	pemrograman	saya
0	2	1	0	1	0	0	0	1	1
1	1	0	0	0	1	0	1	0	2
0	1	0	1	0	0	1	0	0	0

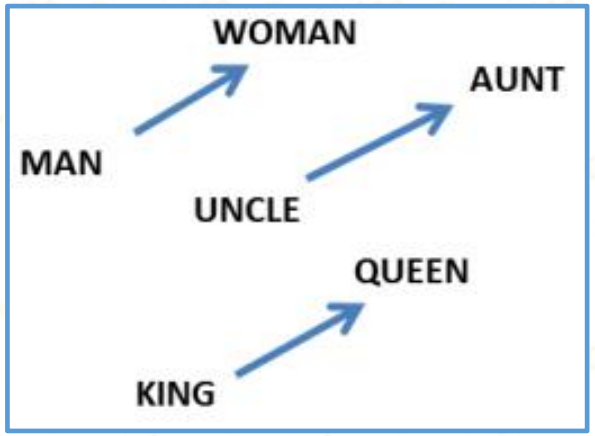
2

adik	belajar	dan	ibu	melukis	membantu	menjahit	menulis	pemrograman	saya
0	2	2.0986123	0	2.0986	0	0	0	2.09861229	1.405465
2.098612	1	0	0	0	2.09861229	0	2.0986123	0	2.81093
0	1	0	2.0986	0	0	2.09861229	0	0	0

Word Embedding

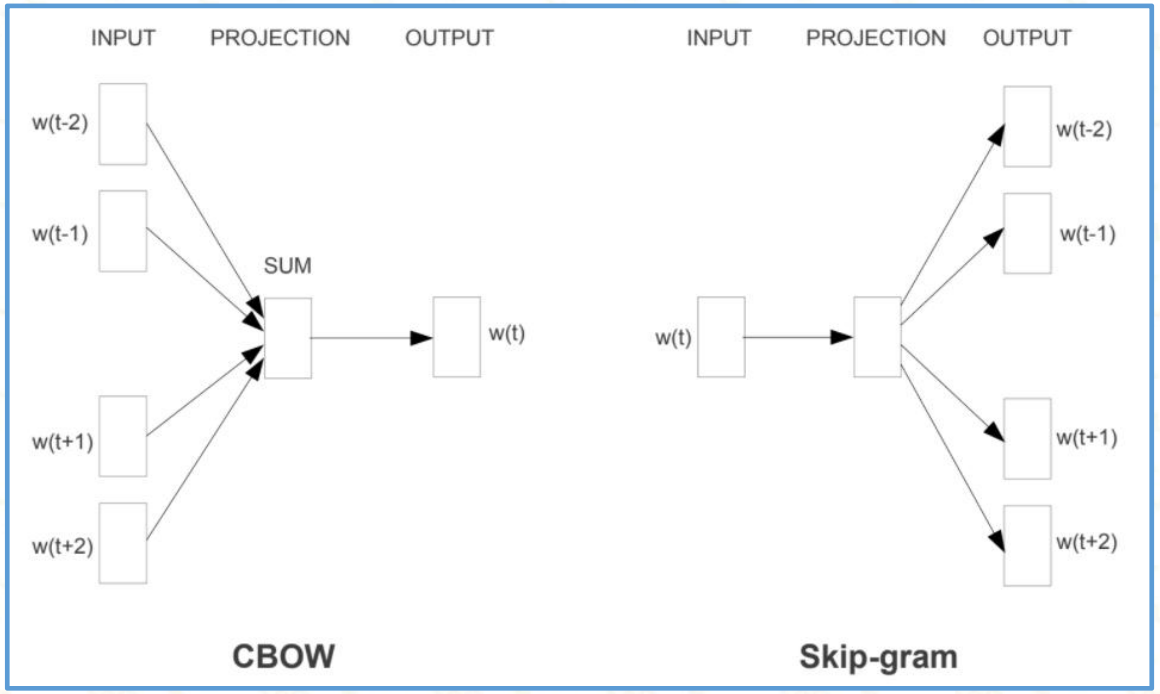
- *Word embedding*
  - Mampu menangkap konteks kata dalam dokumen, kesamaan semantik, hubungan dengan kata lain, dll.
  - Dengan *word embedding*, kata-kata yang memiliki properti tertentu, misalnya berada pada konteks yang sama, atau memiliki *semantic meaning* yang sama berada tidak jauh satu sama lain pada *space* tersebut.





Word2Vec

- Menggunakan model *ANN* untuk mempelajari asosiasi kata dari kumpulan teks yang besar.
- *Word2Vec* terdiri dua teknik yaitu *Continous Bag of Words (CBOW)* dan *Skip Gram Model.*
- The *CBOW* model memprediksi kata saat ini berdasarkan konteksnya.
- *Skip-gram model* belajar dengan memprediksi kata-kata di sekitarnya dengan diberikan kata saat ini.



CBOW model

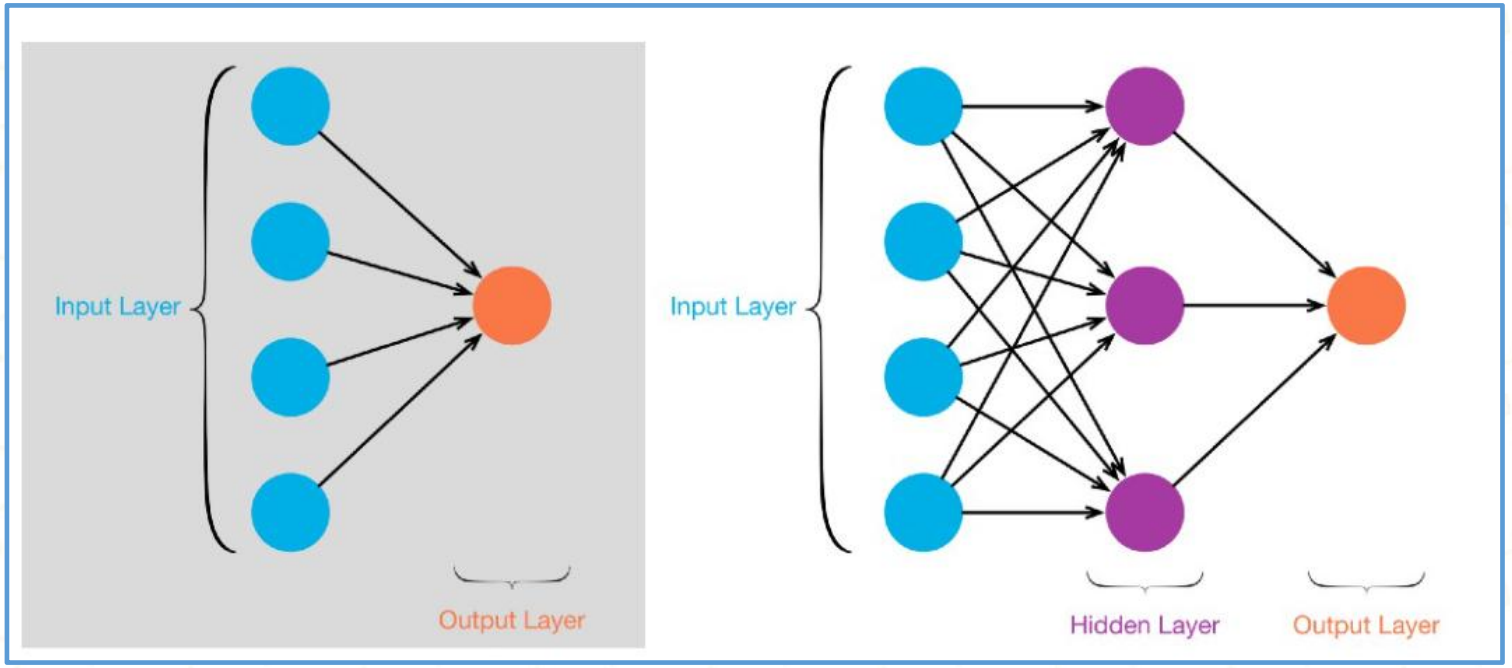
https://machinelearningmastery.com/what-are-word-embeddings/

DigitalSkola
Uncover The World Of Digital Skills With Us

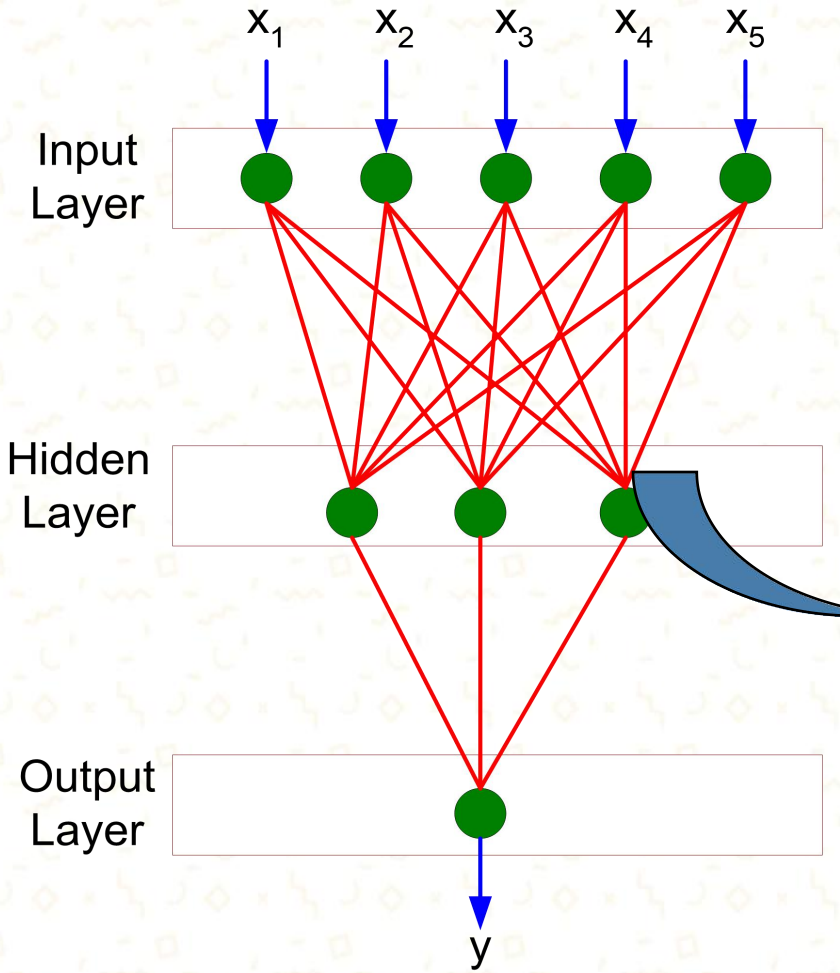
Classification using Machine Learning models

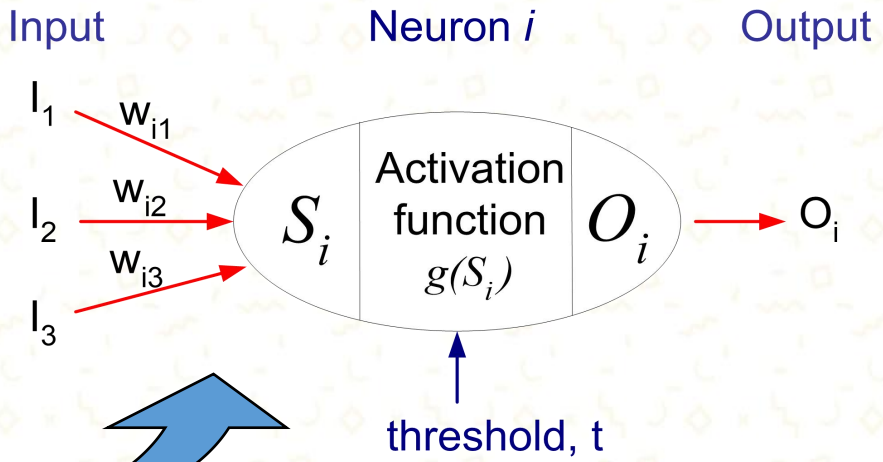
Multilayer perceptron (MLP)

- Type dari Artificial Neural Network Model yang terdiri dari tiga jenis layer — input layer, hidden layer, output layer.
- Kecuali node input, setiap node adalah neuron yang menggunakan fungsi aktivasi nonlinier.
- MLP menggunakan backpropagation untuk training-nya.



General Structure of MLP



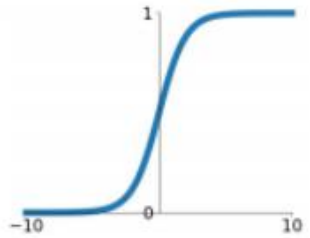


Training ANN means learning the weights of the neurons

Activation function

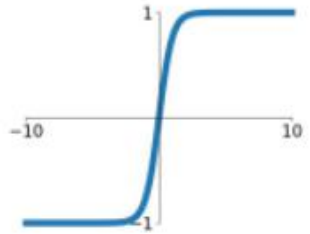
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



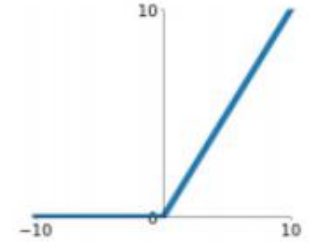
tanh

$$\tanh(x)$$

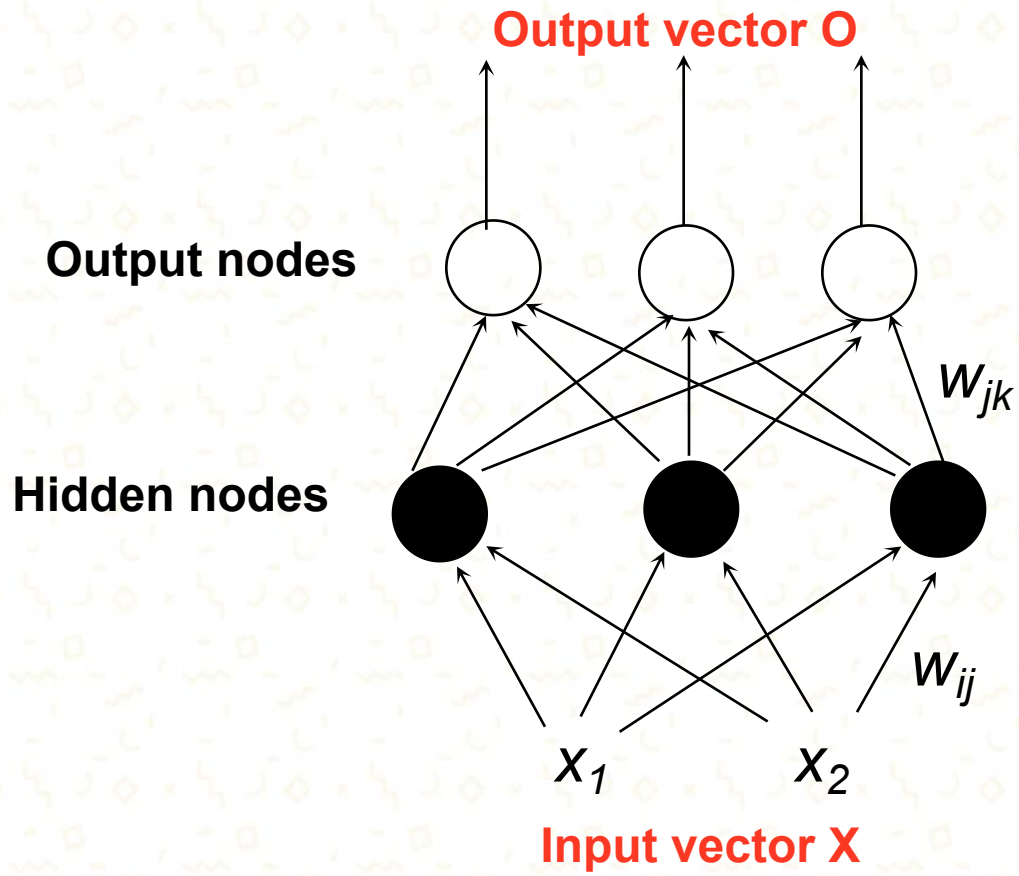


ReLU

$$\max(0, x)$$



Multilayer perceptron

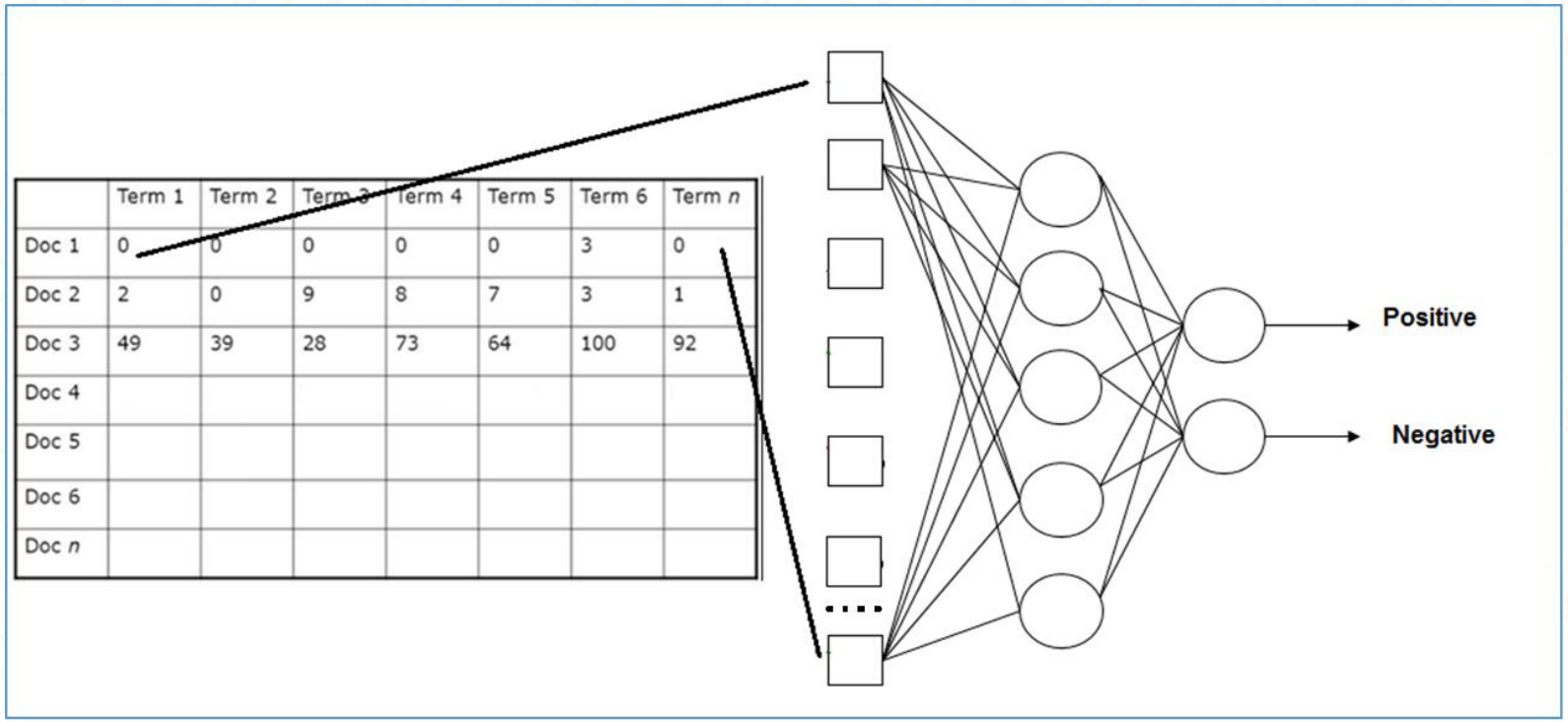


$$O_k = \frac{1}{1+e^{-\sum h_j w_{jk} + \theta_k}}$$

$$h_j = \frac{1}{1+e^{-\sum x_i w_{ij} + \theta_j}}$$

Sentiment analysis using MLP

- Use X_train as features for MLP

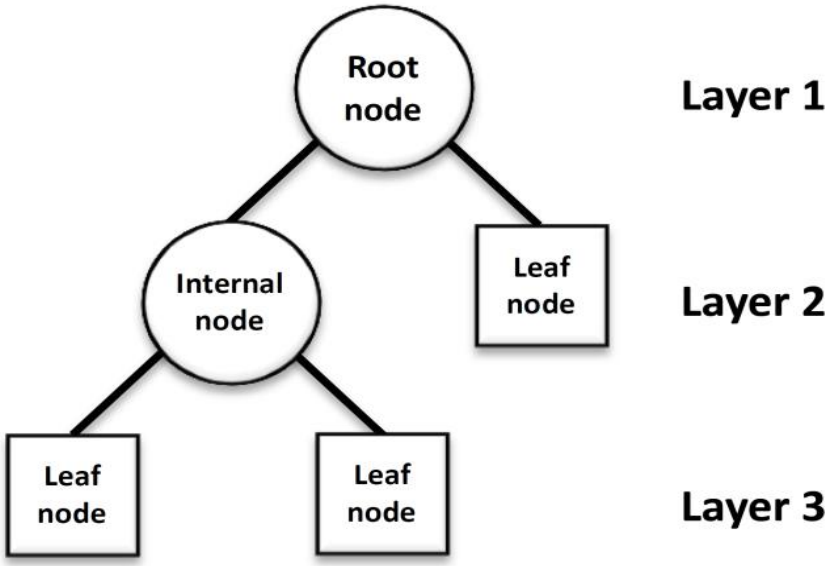


Discussion on NN

- Keuntungan
  - **Robust** -berfungsi baik ketika training set mengandung error
  - Output bisa discrete, real-valued, atau vector
- Kekurangan
  - Waktu yang lama saat training
  - Sulit untuk dipahami

Decision Tree Induction

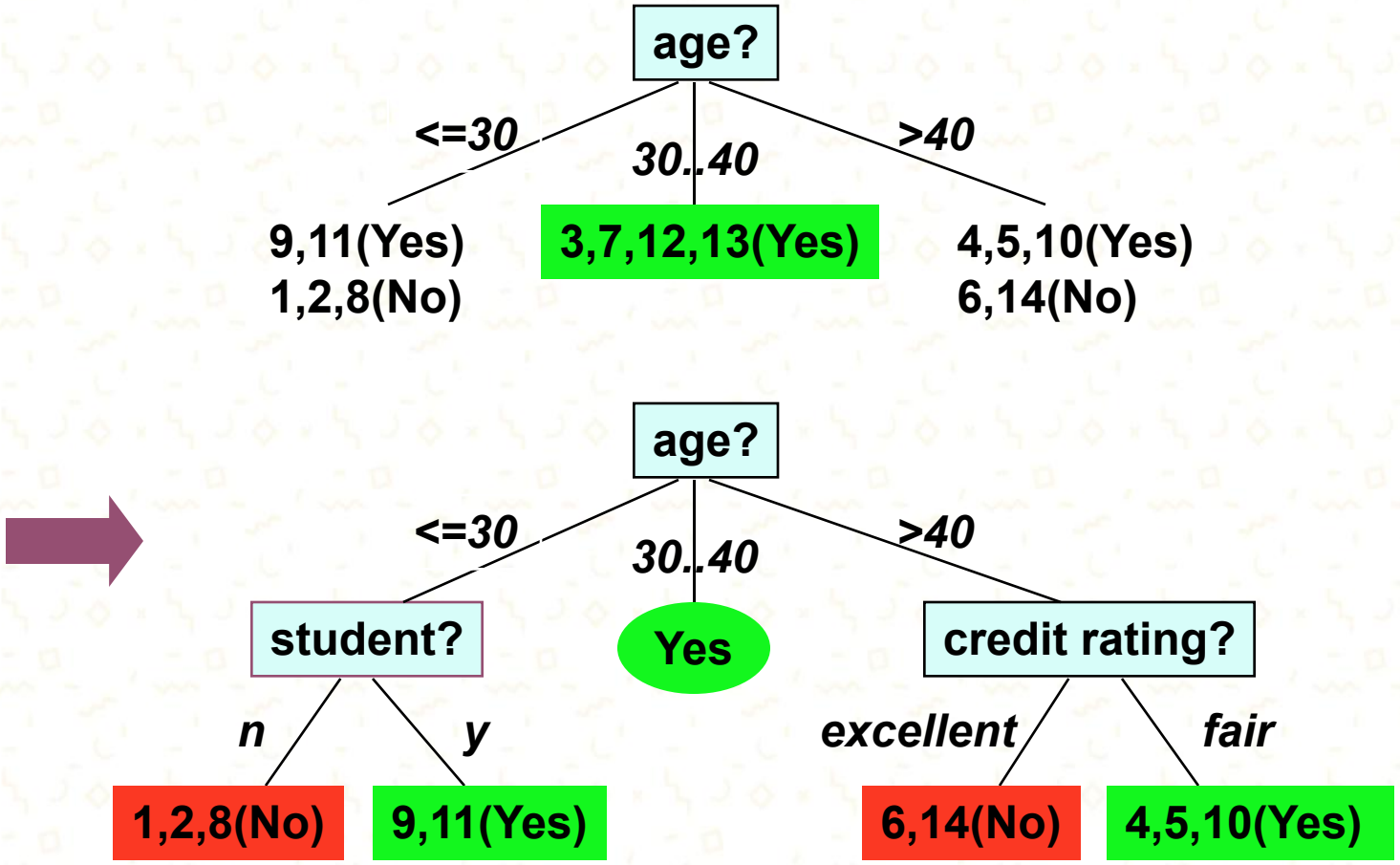
- Basic algorithm
  1. At start, all the training examples are at the root
  2. Test **attributes are selected** on the basis of a heuristic or statistical measure (e.g., information gain)
  3. **Examples are partitioned** recursively based on selected attributes



Training Dataset

No.	age	income	student	credit_rating	buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31…40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31…40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31…40	medium	no	excellent	yes
13	31…40	high	yes	fair	yes
14	>40	medium	no	excellent	no

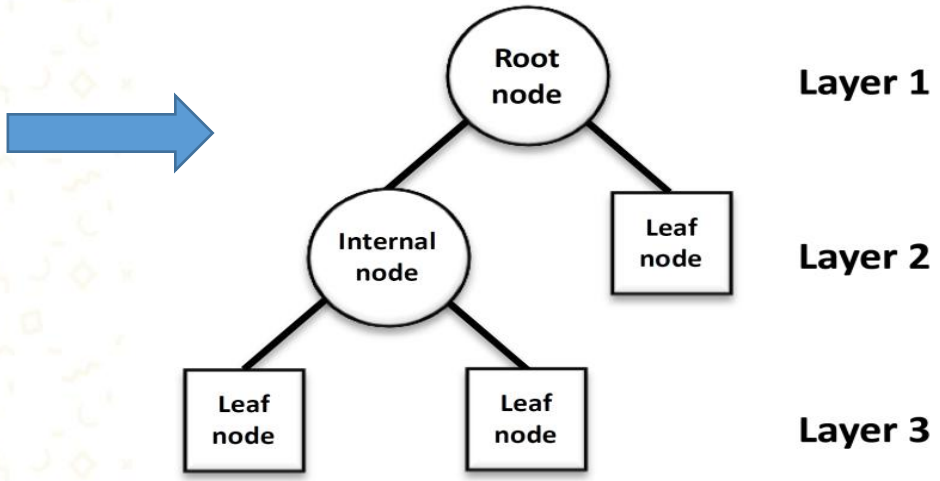
Algorithm for DT Induction - Example



Sentiment analysis using DT

- Use X_train as features for DT

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term *n*	class
Doc 1	0	0	0	0	0	3	0	
Doc 2	2	0	9	8	7	3	1	
Doc 3	49	39	28	73	64	100	92	
Doc 4								
Doc 5								
Doc 6								
Doc *n*								



Discussion on DT

- Kelebihan
  - Dapat diubah menjadi aturan klasifikasi yang dapat dipahami
  - Relatif cepat
- Kekurangan
  - Sensitive (not robust) terhadap noises
  - Continuous-valued attributes - partisi secara dinamis nilai atribut kontinu ke dalam set interval diskrit

DigitalSkola
Uncover The World Of Digital Skills With Us

Let’s practice

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU