Learning Progress Review Week-17

OPTIMISTIC TEAM

Aldiva Wibowo
Asprizal Rizky
Gilang Rahmat
Lutfia Humairosi
Millenia Winadya Putri

Classification for Text Dataset

Definisi Text Classification

Klasifikasi teks adalah proses pemberian tag atau kategori ke teks menurut isinya. Klasifikasi teks dapat digunakan untuk mengatur, menyusun, dan mengkategorikan hampir semua hal. Misalnya, artikel baru dapat diatur berdasarkan topik, percakapan obrolan dapat diatur berdasarkan bahasa, penyebutan merek dapat diatur berdasarkan sentimen, dan sebagainya.

Cara Kerja Text Classification

Ada banyak pendekatan untuk klasifikasi teks otomatis, yang dapat dikelompokkan menjadi tiga jenis sistem yang berbeda :

1. Rule-based systems

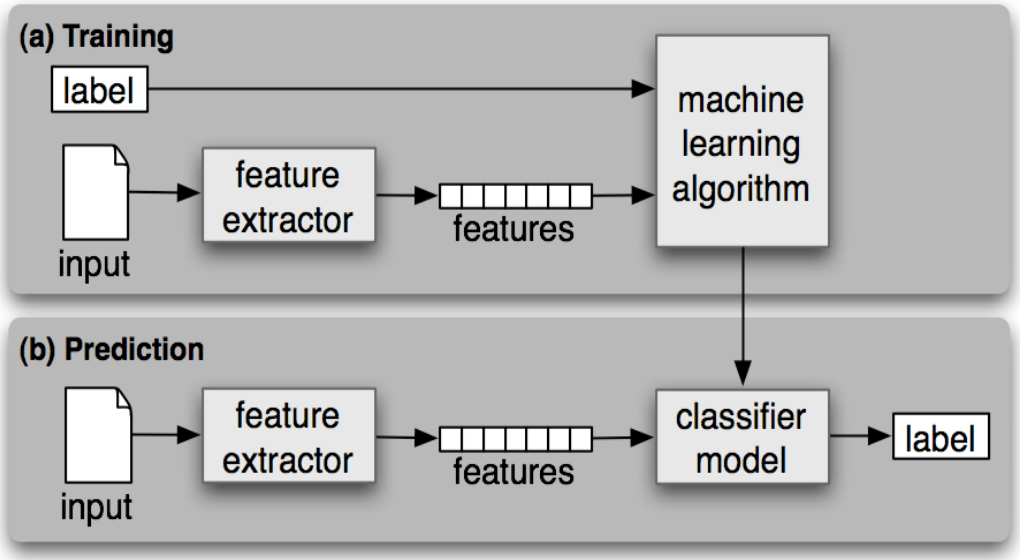
Pendekatan berbasis aturan mengklasifikasikan teks ke dalam kelompok terorganisir dengan menggunakan seperangkat aturan linguistik buatan tangan. Aturan-aturan ini menginstruksikan sistem untuk menggunakan elemen teks yang relevan secara semantik untuk mengidentifikasi kategori yang relevan berdasarkan isinya.

2. Machine Learning based systems

Alih-alih mengandalkan aturan yang dibuat secara manual, klasifikasi teks dengan machine learning membuat klasifikasi berdasarkan pengamatan sebelumnya. Dengan menggunakan contoh yang diberi label sebelumnya sebagai data pelatihan, algoritma machine learning dapat mempelajari asosiasi yang berbeda antara bagian teks dan bahwa keluaran tertentu (yaitu tag) diharapkan untuk masukan tertentu (yaitu teks).

3. Hybrid systems

Sistem hybrid menggabungkan pengklasifikasi dasar yang dilatih dengan machine learning dan rule-based system, yang digunakan untuk lebih meningkatkan hasil. Sistem hybrid ini dapat dengan mudah disesuaikan dengan menambahkan aturan khusus untuk tag yang bentrok yang belum dimodelkan dengan benar oleh pengklasifikasi dasar.



Sentiment Analysis

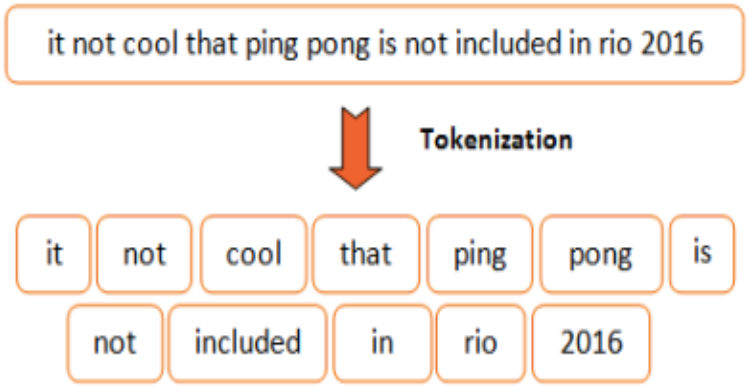
Analisis sentimen adalah proses mendeteksi sentimen positif atau negatif dalam teks. Hal ini sering digunakan oleh bisnis untuk mendeteksi sentimen dalam data sosial, mengukur reputasi merek, dan memahami pelanggan.

Karena pelanggan mengekspresikan pikiran dan perasaan mereka lebih terbuka daripada sebelumnya, analisis sentimen menjadi alat penting untuk memantau dan memahami sentimen itu. Menganalisis umpan balik pelanggan secara otomatis, seperti opini dalam tanggapan survei dan percakapan media sosial, memungkinkan mereka untuk mempelajari apa yang membuat pelanggan senang atau frustrasi, sehingga mereka dapat menyesuaikan produk dan layanan untuk memenuhi kebutuhan pelanggan mereka.

Pre-Processing for Text Classification

DigitalSkola

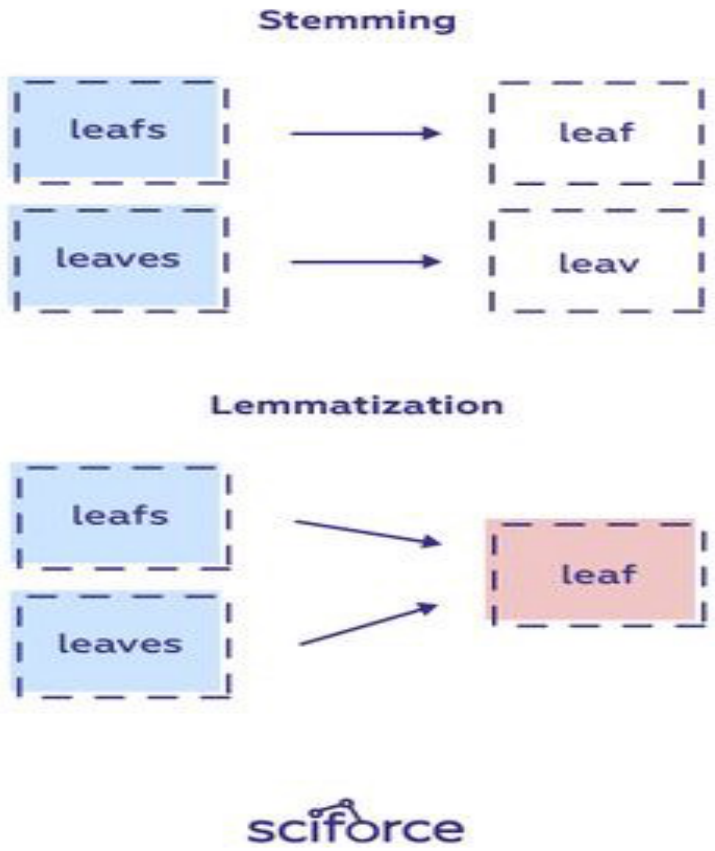
1. Tokenization



2. Removing Stopwords

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

3. Stemming or Lemmatization



Feature Extraction

Review 1: This movie is very scary and long
Review 2: This movie is not scary and is slow
Review 3: This movie is spooky and good

1. Bag of Words

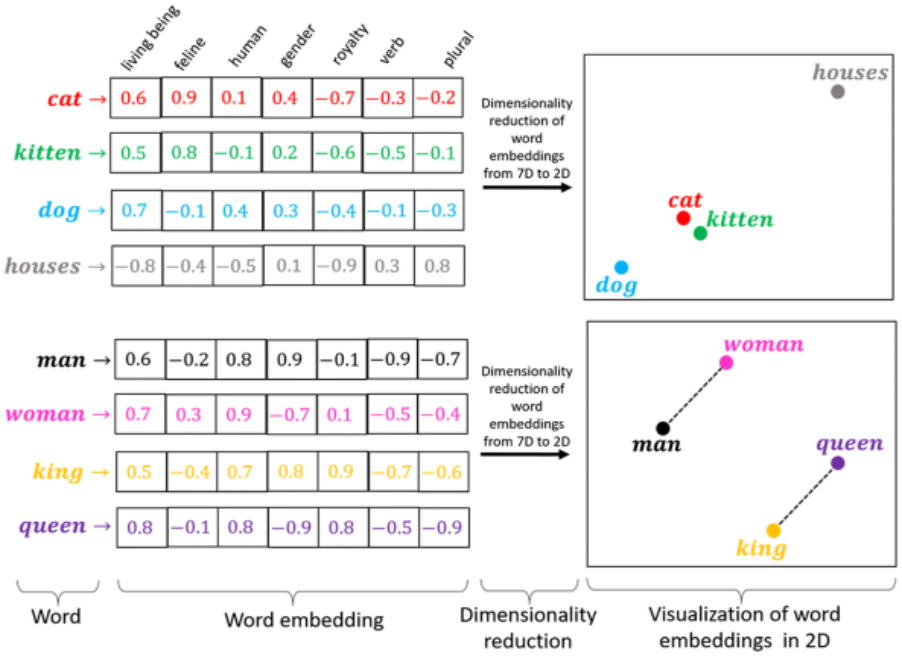
	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

2. TF – IDF (Term Frequency – Inverse Document Frequency)

Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

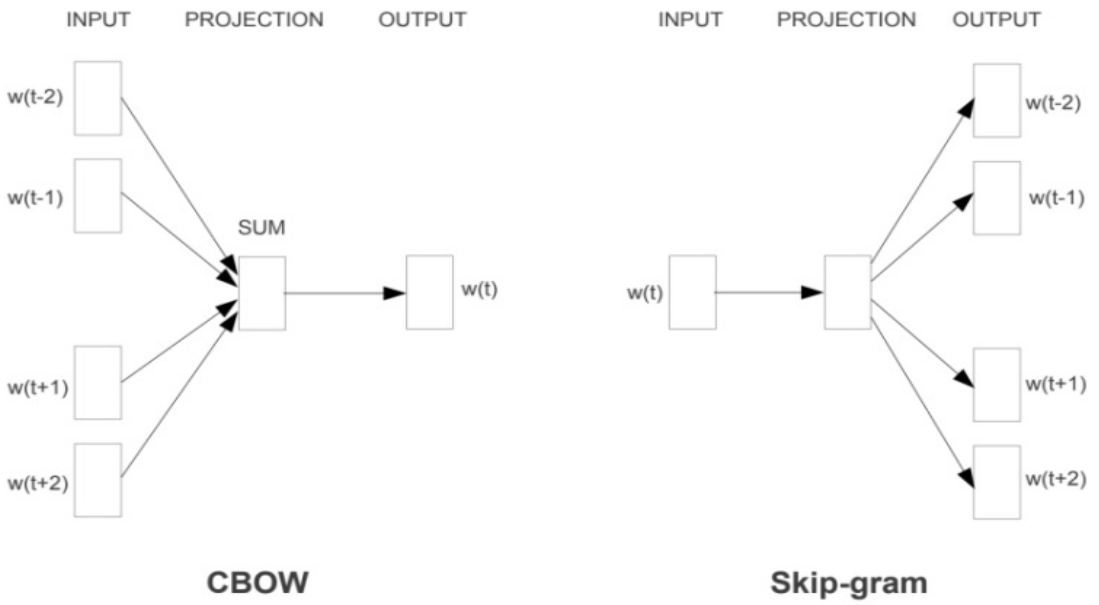
3. Word Embedding

Word embeddings adalah proses konversi kata yang berupa karakter *alphanumeric* kedalam bentuk *vector*. Setiap kata adalah *vector* yang merepresentasikan sebuah titik pada *space* dengan dimensi tertentu. 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.



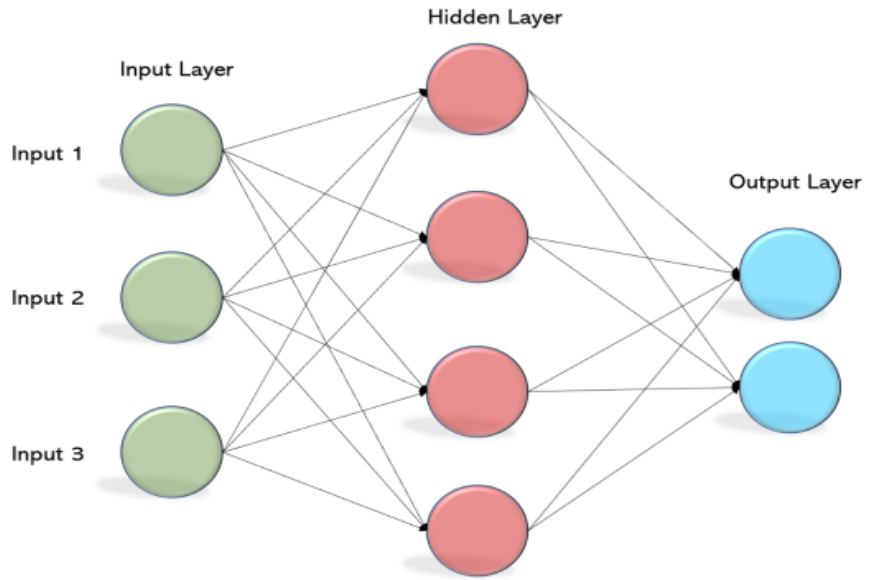
4. Word2Vec

Word2Vec adalah model *shallow neural network* yang merubah representasi kata yang merupakan kombinasi dari karakter *alphanumeric* menjadi *vector*. Representasi *vector* tersebut memiliki properti *relationship* terhadap kata-kata yang berkaitan melalui proses *training*. Terdapat dua model arsitektur yang dapat digunakan pada *word2vec*, yaitu *CBOW* dan *Skip-Gram*



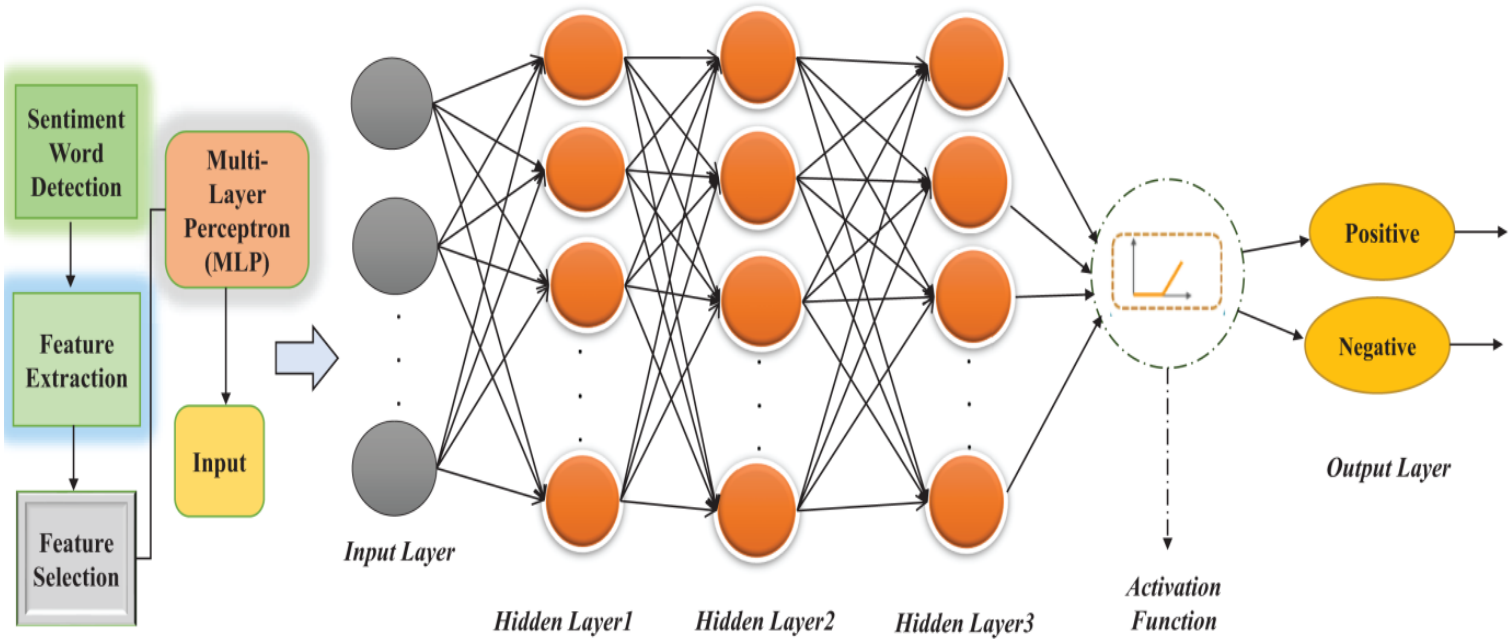
5. MLP (Multilayer Perception)

Multi layer perceptron (MLP), adalah salah satu permodelan dalam teknologi jaringan saraf tiruan (JST) dengan karakteristik memiliki nilai bobot yang lebih baik dari pada pemodelan yang lain, sehingga menghasilkan klasifikasi yang lebih akurat pula. Seperti namanya, multi layer perceptron merupakan pengembangan dari perceptron tunggal, sehingga memiliki beberapa lapisan ataupun hidden layer, yang terletak diantara ruang input dan output layer.



Sentiment Analysis using MLP

DigitalSkola



Discussion on NN

- Keuntungan
  1. Robust -berfungsi baik ketika training set mengandung error
  2. Output bisa discrete, real-valued, atau vector

- Kekurangan
  1. Waktu yang lama saat training
  2. Sulit untuk dipahami

Discussion on DT

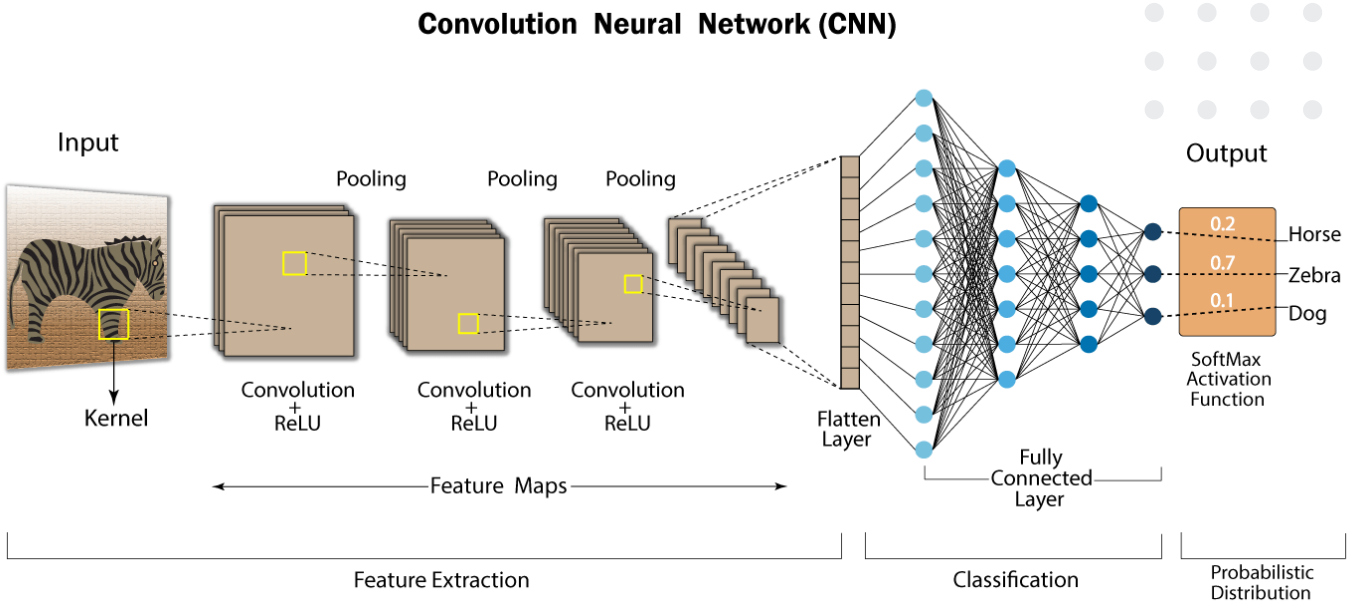
- Kelebihan
  1. Dapat diubah menjadi aturan klasifikasi yang dapat dipahami
  2. Relatif cepat

- Kekurangan
  1. Sensitive (not robust) terhadap noises
  2. Continuous-valued attributes - partisi secara dinamis nilai atribut kontinu ke dalam set interval diskrit

Image Classification

What is Image Classification ?

Klasifikasi gambar adalah dimana komputer dapat menganalisis gambar dan mengidentifikasi 'kelas' gambar tersebut. (Atau probabilitas gambar menjadi bagian dari 'kelas'.) Kelas pada dasarnya adalah label, misalnya, 'mobil', 'binatang', 'bangunan', dan seterusnya.



Why is Image Classification useful?

Why is Image Classification useful?

Klasifikasi gambar memiliki beberapa kegunaan — dan potensi besar seiring dengan meningkatnya keandalannya. Berikut adalah beberapa contoh dari apa yang membuatnya berguna.

Mobil self-driving menggunakan klasifikasi gambar untuk mengidentifikasi apa yang ada di sekitarnya. Yaitu. pohon, orang, lampu lalu lintas dan sebagainya.

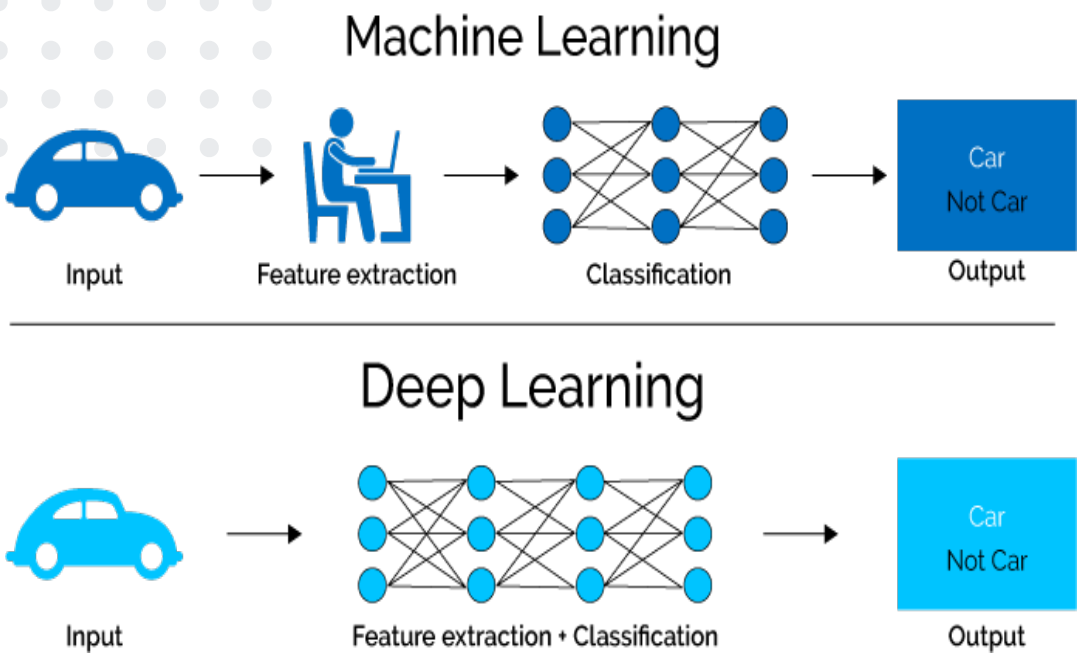
Klasifikasi gambar juga dapat membantu dalam perawatan kesehatan. Misalnya, dapat menganalisis gambar medis dan menyarankan apakah gambar tersebut diklasifikasikan sebagai menggambarkan gejala penyakit.

Challenge in Image Classification

Klasifikasi citra awal mengandalkan data piksel mentah. Ini berarti bahwa komputer akan memecah gambar menjadi piksel individu. Masalahnya adalah dua gambar dari hal yang sama bisa terlihat sangat berbeda. Mereka dapat memiliki latar belakang, sudut, pose, dan lain-lain yang berbeda. Ini membuatnya menjadi tantangan bagi komputer untuk 'melihat' dan mengkategorikan gambar dengan benar. Dan saat ini sudah mengalami perkembangan untuk klasifikasi gambar dapat menggunakan CNN (Convolutional Neural Network).

Introduction to Deep Learning

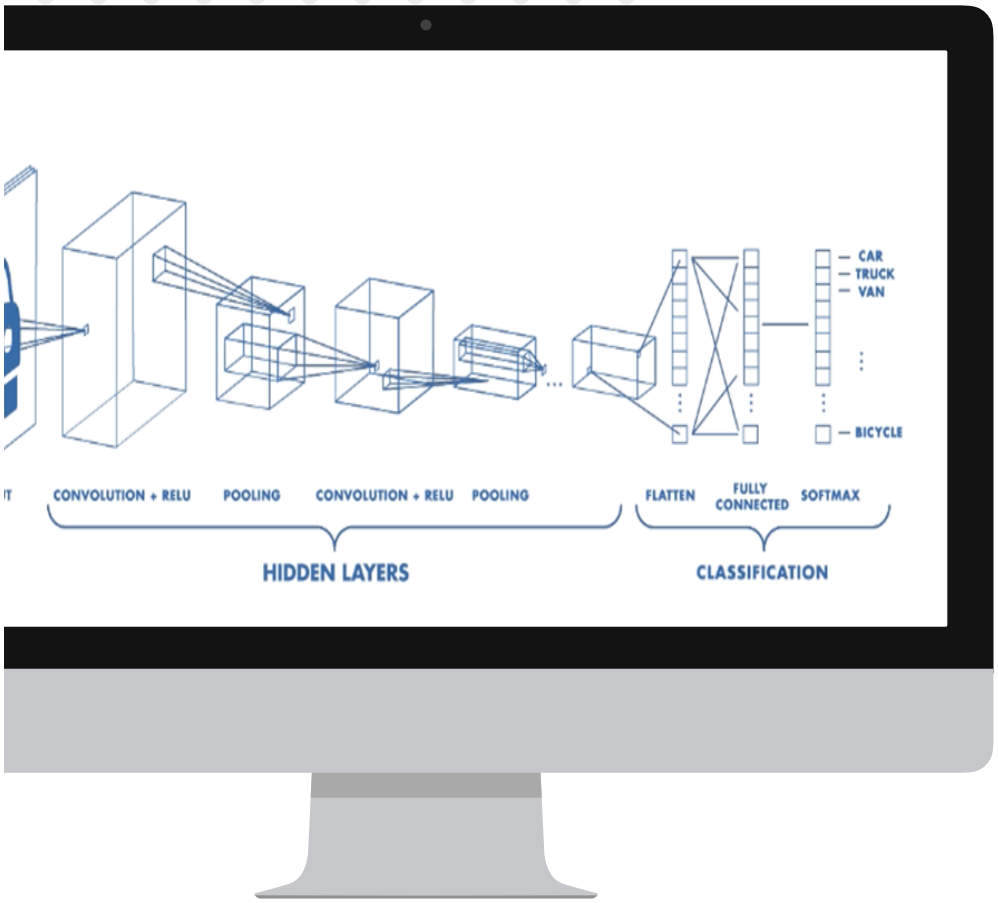
Pendekatan klasifikasi secara konvensional umumnya melakukan ektraksi fitur secara terpisah kemudian dilanjutkan proses pembelajaran menggunakan metode klasifikasi konvensional.



Kelemahan pendekatan konvensional:

- Memerlukan waktu dan pengetahuan lebih untuk ekstraksi fitur
- Sangat tergantung pada satu domain permasalahan saja sehingga tidak berlaku general

Pendekatan klasifikasi berbasis Deep learning mempelajari representasi hirarki (pola fitur) secara otomatis melalui beberapa tahapan proses feature learning



Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) adalah jenis jaringan saraf tiruan yang digunakan dalam pengenalan dan pemrosesan gambar yang dirancang khusus untuk memproses data piksel. CNN adalah pemrosesan gambar yang kuat, kecerdasan buatan (AI) yang menggunakan pembelajaran mendalam untuk melakukan tugas generatif dan deskriptif.

- CNN merupakan metode Deep Learning yang merupakan salah satu jenis arsitektur ANN

- Ada tiga layer utama yaitu convolutional layer, pooling layer, dan fully connected layer

Inventor of Convolutional Neural Network

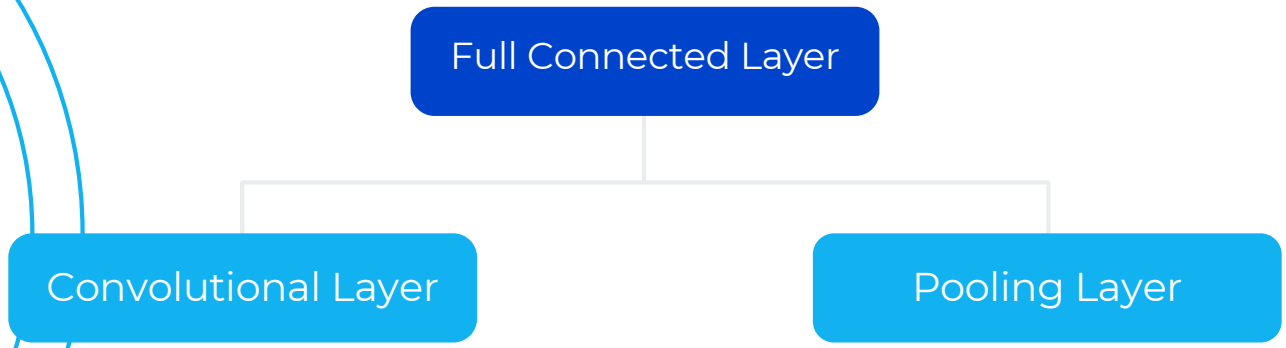
Yann André LeCun adalah seorang ilmuwan komputer Prancis bekerja di bidang Machine Learning, Computer Vision, Mobile Robotics, dan Computational Neuroscience. Dia adalah salah satu murid Geoffrey Hinton (Geoffrey adalah Godfather of Deep learning)

CNN Architecture Section

Convolutional adalah cara matematis menggabungkan dua sinyal untuk membentuk sinyal ketiga. Dalam pemrosesan gambar, konvolusi berhubungan dengan penerapan filter ke gambar. Konvolusi mencakup membalik dan mengalikan, tetapi karena filter umumnya simetris, pemfilteran gambar umum hanya mencakup perkalian.

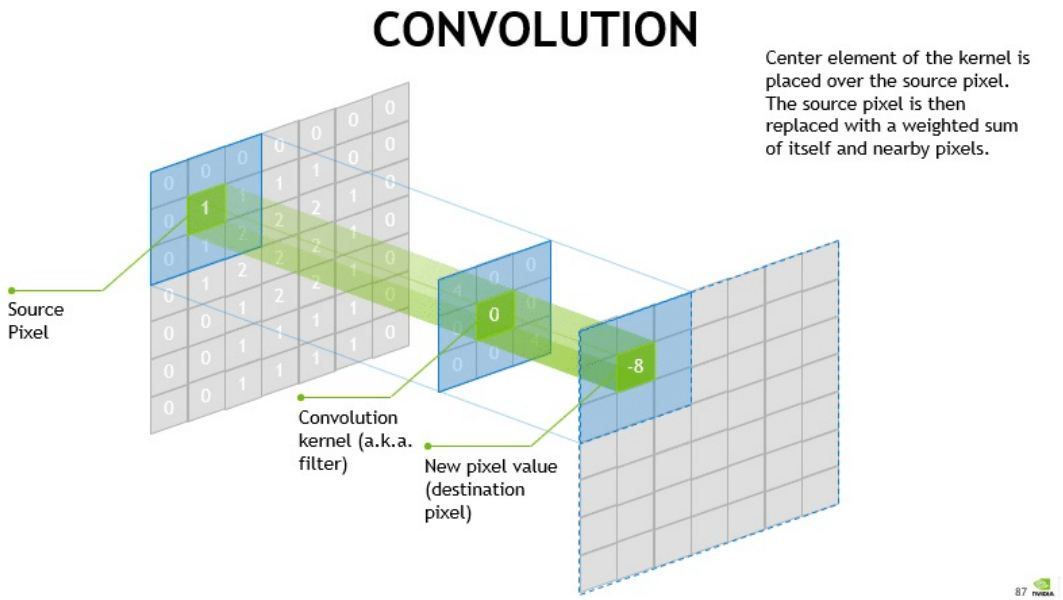
DigitalSkola

CNN Architecture



Convolutional Layer

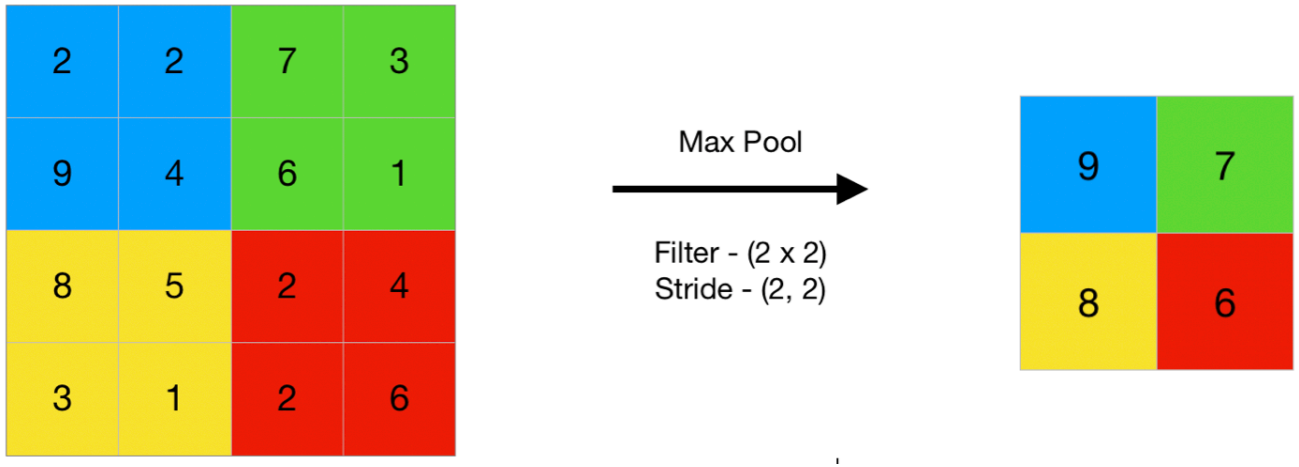
Menggunakan operasi konvolusi dari teori pengolahan citra. Berperan untuk menghasilkan “feature image/map,”gambar yang berisi fitur penting dari gambar input. Ukuran matrik citra dan ukuran matrik filter akan mempengaruhi ukuran matrik feature map.



Pooling Layer

Berperan untuk memperkecil dimensi feature image layer yang berperan untuk mereduksi dimensionalitas output dari layer sebelumnya. Membuat ukuran feature image menjadi lebih kecil.

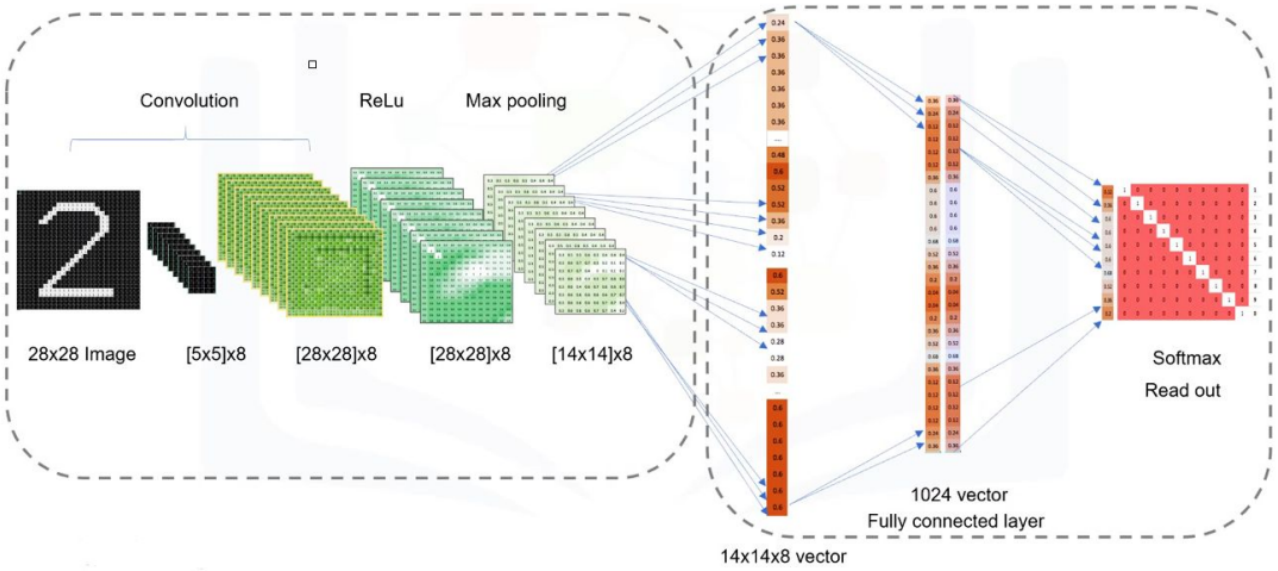
Jenis: Max-pooling, Average pooling, dll.



Fully-connected Layer

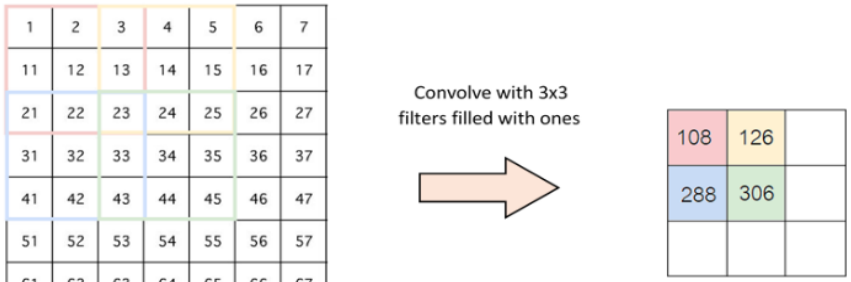
Multi Layer Perceptron biasa Berperan untuk menghasilkan output klasifikasi akhir. Feature map hasil dari proses konvolusi dan pooling, selanjutnya dilakukan proses flatten yaitu merubah matrix menjadi vektor sebagai inputan fully connected layer.

Disetiap layer fully-connected activation function yang digunakan bebas, Kecuali akhir layer fully-connected. Layer akhir bertugas memberikan probabilitas klasifikasi dengan Softmax Function.

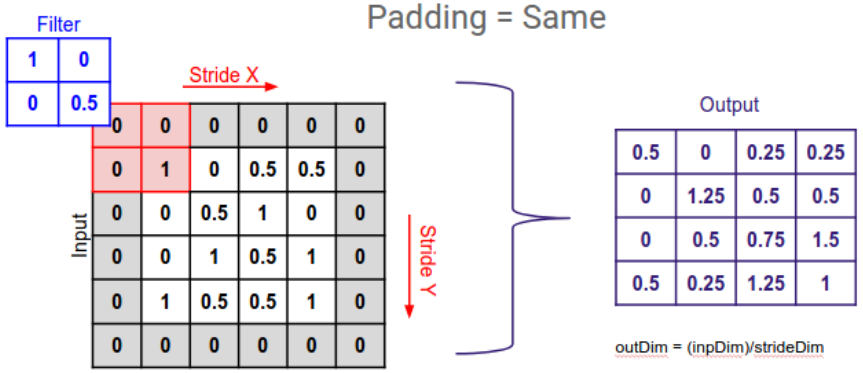


More in CNN

Stride
adalah jumlah piksel yang bergeser di atas matriks input. Ketika langkahnya adalah 1 maka kami memindahkan filter ke 1 piksel sekaligus. Ketika langkahnya adalah 2 maka kami memindahkan filter ke 2 piksel sekaligus dan seterusnya.



Padding
adalah istilah yang relevan dengan convolutional neural network karena mengacu pada jumlah piksel yang ditambahkan ke gambar ketika sedang diproses oleh kernel CNN. Misalnya, jika padding di CNN disetel ke nol, maka setiap nilai piksel yang ditambahkan akan bernilai nol.



Example

```python
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D, Flatten, Dense

# Menggunakan Convolutional Neural Network

model2 = Sequential()
model2.add(Conv2D(16,(3,3),activation='relu',input_shape=(28,28,1),padding='same'))
model2.add(MaxPooling2D(2,2))
model2.add(Conv2D(32,(3,3),activation='relu',padding='same'))
model2.add(MaxPooling2D(2,2))

model2.add(Flatten())
model2.add(Dense(64,activation='relu'))
model2.add(Dense(10,activation='softmax'))
```

executed in 69ms, finished 19:35:16 2021-12-16

```python
model2.summary()
```

executed in 12ms, finished 19:35:19 2021-12-16

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 28, 28, 16)        160

 max_pooling2d (MaxPooling2D  (None, 14, 14, 16)       0
 )

 conv2d_1 (Conv2D)           (None, 14, 14, 32)        4640

 max_pooling2d_1 (MaxPooling  (None, 7, 7, 32)         0
 2D)

 flatten_2 (Flatten)         (None, 1568)              0

 dense_4 (Dense)             (None, 64)                100416

 dense_5 (Dense)             (None, 10)                650

=================================================================
Total params: 105,866
Trainable params: 105,866
Non-trainable params: 0
_________________________________________________________________
```

```python
model2.compile(optimizer='adam',loss='categorical_crossentropy',metrics=['acc'])

history2 = model2.fit(X_train,y_train,epochs=10,batch_size=100,validation_data=(X_test,y_test))
```

executed in 3m 37s, finished 19:39:01 2021-12-16

```
Epoch 1/10
600/600 [==============================] - 22s 35ms/step - loss: 0.2405 - acc: 0.9296 - val_loss: 0.0661 - val_acc: 0.9792
Epoch 2/10
600/600 [==============================] - 22s 37ms/step - loss: 0.0654 - acc: 0.9799 - val_loss: 0.0438 - val_acc: 0.9859
Epoch 3/10
600/600 [==============================] - 22s 37ms/step - loss: 0.0497 - acc: 0.9847 - val_loss: 0.0404 - val_acc: 0.9855
Epoch 4/10
600/600 [==============================] - 23s 38ms/step - loss: 0.0375 - acc: 0.9882 - val_loss: 0.0353 - val_acc: 0.9878
Epoch 5/10
600/600 [==============================] - 21s 35ms/step - loss: 0.0308 - acc: 0.9905 - val_loss: 0.0364 - val_acc: 0.9867
Epoch 6/10
600/600 [==============================] - 21s 36ms/step - loss: 0.0259 - acc: 0.9920 - val_loss: 0.0366 - val_acc: 0.9875
Epoch 7/10
600/600 [==============================] - 21s 35ms/step - loss: 0.0209 - acc: 0.9934 - val_loss: 0.0326 - val_acc: 0.9888
Epoch 8/10
600/600 [==============================] - 21s 35ms/step - loss: 0.0190 - acc: 0.9937 - val_loss: 0.0392 - val_acc: 0.9874
Epoch 9/10
600/600 [==============================] - 21s 35ms/step - loss: 0.0156 - acc: 0.9948 - val_loss: 0.0352 - val_acc: 0.9890
Epoch 10/10
600/600 [==============================] - 22s 36ms/step - loss: 0.0117 - acc: 0.9963 - val_loss: 0.0349 - val_acc: 0.9900
```

```python
model2.evaluate(X_test,y_test)
```

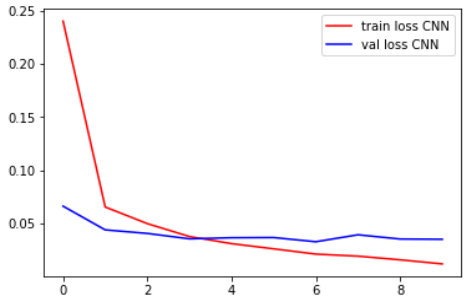
executed in 2.01s, finished 19:39:06 2021-12-16

```
313/313 [==============================] - 2s 5ms/step - loss: 0.0349 - acc: 0.9900
```

```python
import matplotlib.pyplot as plt

epochs = range(10)

loss2 = history2.history['loss']
val_loss2 = history2.history['val_loss']

plt.plot(epochs,loss2,'r',label='train loss CNN')
plt.plot(epochs,val_loss2,'b',label='val loss CNN')
plt.legend()
```

executed in 150ms, finished 19:39:08 2021-12-16

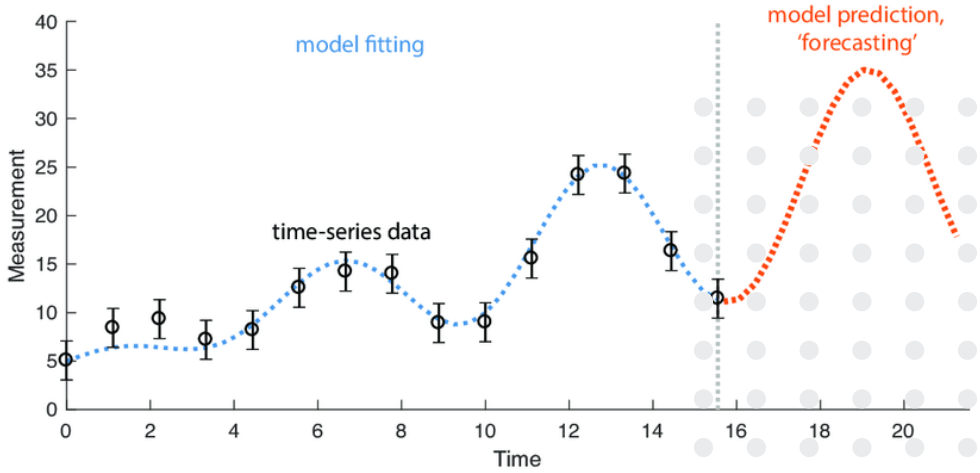
```
<matplotlib.legend.Legend at 0x1e0babfa430>
```



Time Series Forecasting

Time Series Forecasting

Forecasting adalah suatu teknik untuk memprediksi kejadian yang akan datang, dengan menganalisa tren-tren dimasa lalu.



Time Series Forecasting

Adapun tujuan dari forecasting sebagai berikut:

- Untuk mengkaji kebijakan perusahaan yang berlaku saat ini dan dimasa lalu serta melihat sejauh mana pengaruh dimasa datang.
- Peramalan diperlukan karena adanya time lag atau delay antara saat suatu kebijakan perusahaan ditetapkan dengan saat implementasi.
- Peramalan merupakan dasar penyusutan bisnis pada suatu perusahaan sehingga dapat meningkatkan efektivitas suatu rencana bisnis.

Contoh penggunaan Forecasting berdasarkan fungsi dan tujuan :

1. General business forecasting, peramalan bisnis secara keseluruhan mulai dari ekonomi, politik, sosial, budaya dan hal-hal lainnya yang bersifat makro.

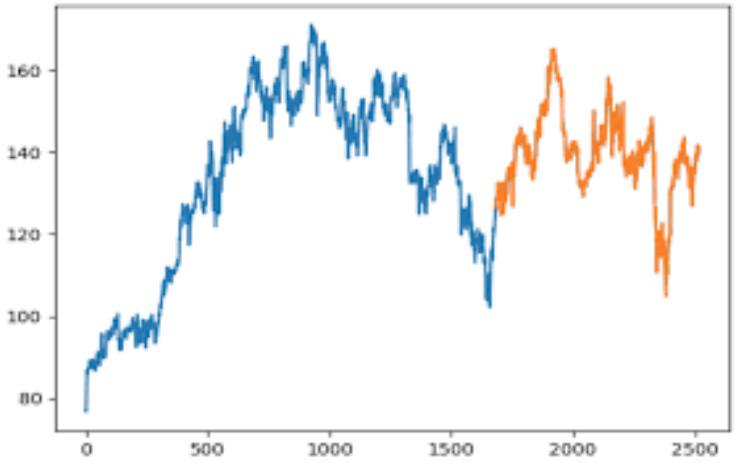
2. Sales forecasting, peramalan jumlah barang yang bisa dijual di masa mendatang berdasarkan data penjualan sebelumnya.

3. Demand forecasting, peramalan yang bertujuan untuk mengetahui perkiraan permintaan dan kondisi pasar.

4. Financial forecasting, atau biasa disebut juga dengan capital forecasting. Bertujuan untuk memperkirakan biaya dan modal yang dikeluarkan di masa mendatang.

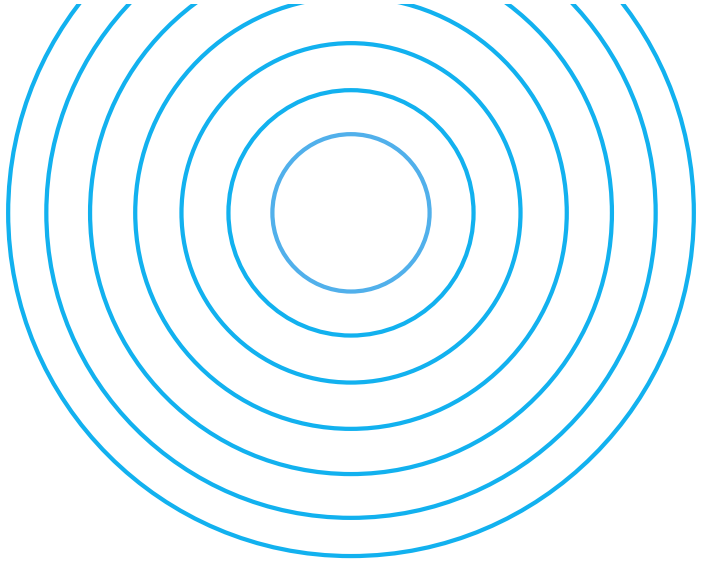
Time Series

Menurut Wikipedia, data Time Series adalah serial dari kumpulan data yang teratur oleh urutan waktu. Frekuensi urutan waktu yang dimiliki oleh Time series data bisa meliputi tahunan, bulanan, jam, atau bahkan mili-detik.



Komponen Data Time series

1. Base atau Level, nilai dari data jika serial data tersebut merupakan garis lurus.
2. Trend, kemiringan meningkat atau menurun yang terlihat di Time Series. Di dalam Trend juga terdapat komponen Cyclic yang berbeda dari Trend tetapi sering digabungkan menjadi satu dengan Trend.
3. Seasonality, pola unik yang terlihat di suatu interval waktu karena faktor musiman. Hal ini bisa karena suatu bulan di setiap tahunnya, suatu hari di setiap bulannya, atau bahkan suatu jam di dalam satu hari.
4. Residual, variasi dari data yang tidak dapat dijelaskan.



Statistical and Machine Learning Approach

Statistical Model

1. Simple Moving Average

Simple moving average adalah bentuk paling sederhana dari moving average. SMA dihitung dengan rumus sebagai berikut:

SMA = (A1 + A2 + ...An) / n

A adalah nilai rata-rata di n. Sementara, n sendiri adalah jumlah periode waktu.

1. Exponential Smoothing

adalah suatu tipe teknik peramalan rata-rata bergerak yang melakukan penimbangan terhadap data masa lalu dengan cara eksponensial sehingga data paling akhir mempunyai bobot atau timbangan lebih besar dalam rata-rata bergerak.

1. Autoregressive Integrated Moving Average

ARIMA merupakan gabungan dari AR dan MA dimana AR adalah singkatan dri autoregresif dan MA merupakan moving average sedangkan I yang ditengah merupakan integrated dimana kegunaannya untuk differensiasi jika data tidak stasioner.

Statistical Model

1. Simple Moving Average

```
# Simple Moving Average
df_sma = data.copy()
df_sma['6-month-SMA'] = data['Passengers'].rolling(window=6).mean()
df_sma['12-month-SMA'] = data['Passengers'].rolling(window=12).mean()
df_sma.head(20)
```

2. Exponential Smoothing

```
# Exponential Smoothing
from statsmodels.tsa.holtwinters import SimpleExpSmoothing

df_ses = data.copy()

model_ses = SimpleExpSmoothing(data['Passengers'])
fitted_model_ses = model_ses.fit(smoothing_level=0.3,optimized=False,use_brute=True)
df_ses['SES'] = fitted_model_ses.fittedvalues

df_ses.head(20)
```

3. Autoregressive Integrated Moving Average

```
from statsmodels.tsa.arima_model import ARIMA
model=ARIMA(train['Passengers'],order=(1,1,1))
results=model.fit()
```

Machine Learning Model

1. Linear Regression
regresi linear merupakan pendekatan untuk memodelkan hubungan antara suatu (satu atau lebih) variabel dependen dengan satu (regresi linear sederhana) atau lebih variabel independen (regresi linier banyak).

2. Random Forest
Random forest adalah suatu algoritma yang digunakan untuk klasifikasi data dalam jumlah yang besar. Random forest merupakan kombinasi dari masing – masing pohon (tree) dari model Decision Tree yang baik, dan kemudian dikombinasikan ke dalam satu model.

3. Long Short Term Memory
Long Short Term Memory (LSTM) merupakan salah satu pengembangan neural network yang dapat digunakan untuk pemodelan data time series [10]. LSTM mampu mengatasi ketergantungan jangka panjang (long term dependencies) pada masukannya.