Text Mining

Document **Classification**

| Outline

1. Text Classification Overview
2. Metode Text Classification
3. Studi Kasus Text Classification dengan KNN
4. Evaluasi

| #1 Text Classification Overview

Classification Learning (a.k.a supervised learning)

- Diberikan sekumpulan data dari obyek yang sudah terlabeli (disebut *training examples*)
- Kemudian, berikan label untuk data baru yang belum pernah terlihat sebelumnya
  - Contoh: Diberikan sekumpulan sample dari transaksi kartu kredit yang sudah diberikan label “penipuan” dan “bukan-penipuan”, kemudian **berikan label untuk data transaksi baru**, apakah transaksi tersebut termasuk penipuan atau bukan!
- Apa bedanya dengan *Clustering*?

| #1 Text Classification Overview

Apakah ini spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

© Stanford University. All Rights Reserved.

| #1 Text Classification Overview

Review film, positif atau negatif?

- Sangat mengecewakan

- Banyak karakter lucunya dan satir-satir yang ada di film ini diimplementasikan dengan sangat baik.

- Ini adalah film komedi terbaik yang pernah ada

- Sangat mengecewakan, bagian yang paling jelek adalah adegan tinju-tinjuannya.

| #1 Text Classification Overview

Tujuan Klasifikasi Teks → Memberikan label untuk setiap dokumen

- Label biasanya berupa **topik** seperti Yahoo-categories
  - contoh: “finansial," “olahraga"
- Label juga bisa berupa **genre**
  - contoh: "editorial" "review-film" “berita“
- Label juga bisa berupa **opini**
  - contoh: “suka”, “benci”, “netral”
- Label juga bisa berupa **pernyataan binary**
  - contoh: “menarik" : “tidak menarik”
  - contoh: “spam” : “bukan spam”
  - contoh: “mengandung bahasa fulgar” “tidak mengandung bahasa fulgar”

| #1 Text Classification Overview

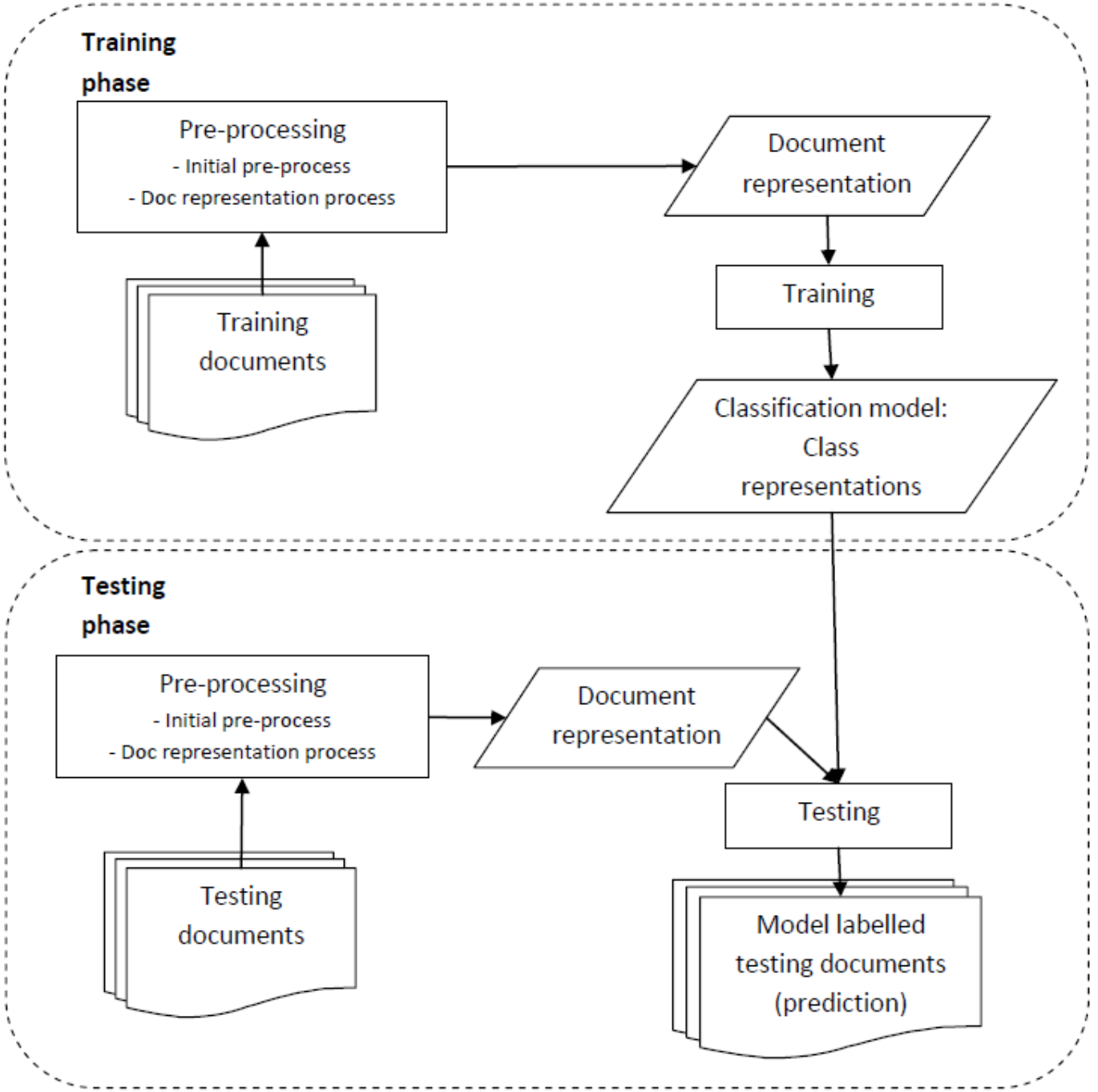
Proses Klasifikasi

1. Manusia memberikan label berupa kelas yang sudah didefinisikan pada obyek-obyek yang ada (data training)

2. Mesin “belajar” dari sample data untuk menghasilkan model statistik

3. Mesin memprediksi kelas dari data baru menggunakan model statistik yang sudah didapatkan pada proses sebelumnya

| #1 Text Classification Overview



| #1 Text Classification Overview

Klasifikasi Teks: Definisi

- *Input*:
  - sebuah dokumen d
  - himpunan kelas $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: prediksi kelas $c \in C$

| #1 Text Classification Overview

Klasifikasi Teks: Pembentukan Data Training

Proses:

- User mengidentifikasi himpunan kelas $C = \{c_1, c_2, \dots, c_n\}$
- User mencari dokumen-dokumen untuk membentuk himpunan dokumen training T
- User memberikan label untuk setiap dokumen $d \, \epsilon \, \mathrm{T}$ menggunakan kelasnya masing-masing
- Output → himpunan data training T_c untuk setiap kelas *c*

| #1 Text Classification Overview

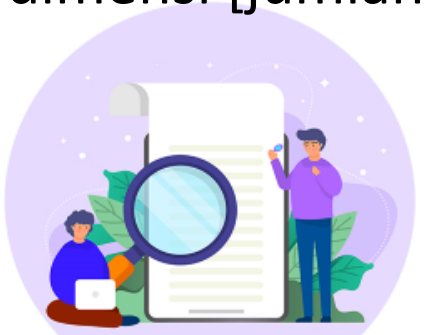
Klasifikasi Teks: Pembentukan Fitur

Membentuk representasi dari obyek yang akan diklasifikasikan. Representasi ini harus *calculable* (dapat dihitung)

Proses:

- Identifikasi himpunan dari fitur-fitur diskrit
- Setiap dokumen direpresentasikan dengan sebuah vektor fitur
- Output → sebuah matriks berdimensi [jumlah object × *jumlah* feature]

→ menggunakan TF-IDF



| #2 Metode Text Classification

Metode Klasifikasi: Hand-coded rules

→ Metode klasifikasi yang berdasarkan pada aturan-aturan yang tertulis.

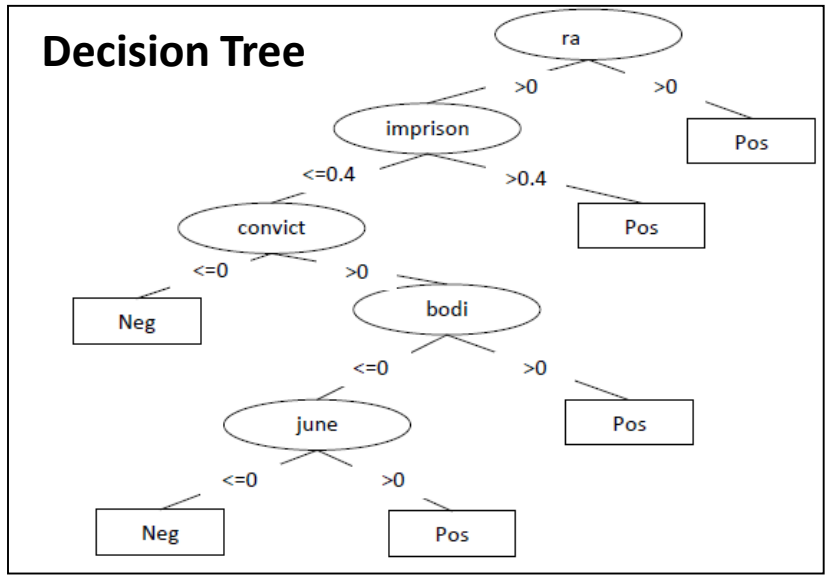
- Aturan didefinisikan berdasarkan kombinasi dari kata atau fitur-fitur lain.
  - spam: alamat-email-masuk-blacklist OR (“juta” AND “selamat”)
- Metode ini bisa menghasilkan akurasi yang tinggi
  - JIKA aturan-aturannya secara hati-hati didefinisikan dan diperbaiki oleh orang yang ahli (expert)
- Tetapi, membangun dan merawat aturan-aturan ini membutuhkan resource yang banyak dan mahal.

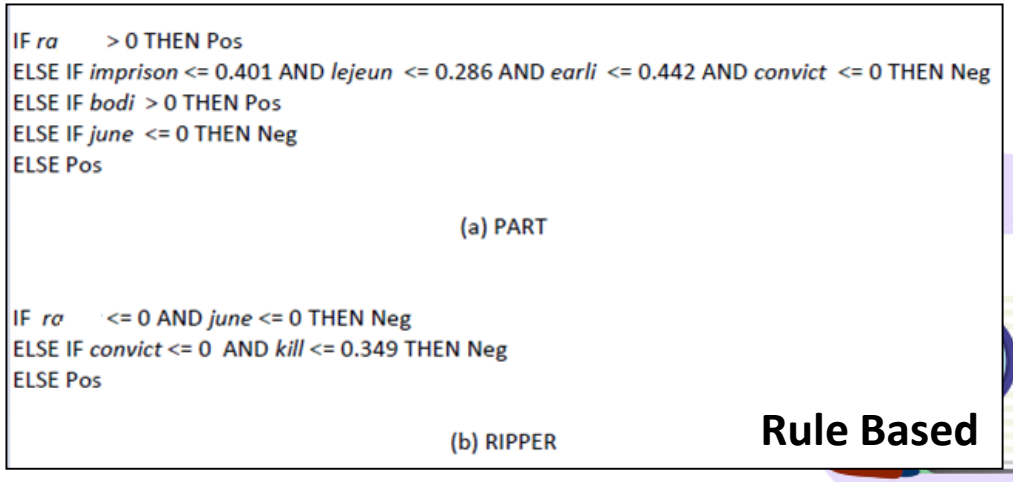
| #2 Metode Text Classification

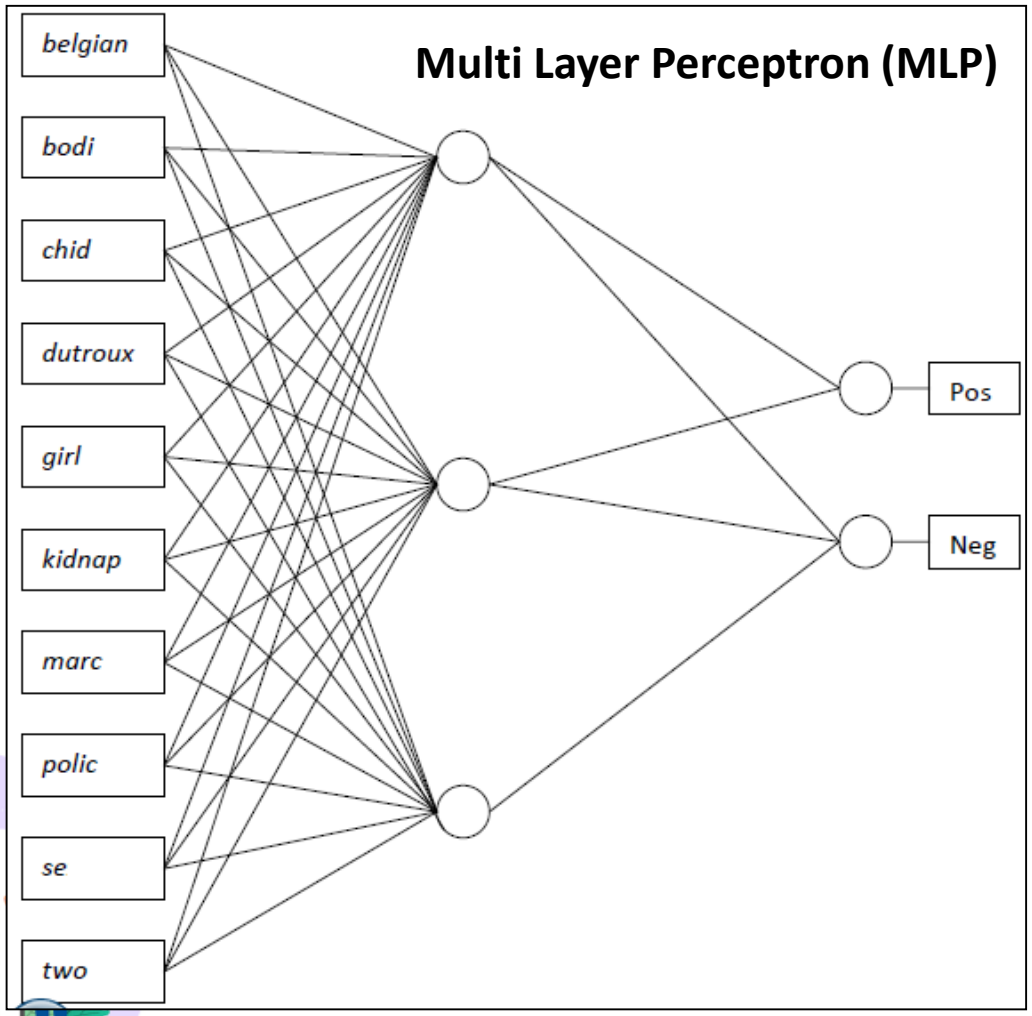
Metode Klasifikasi: Supervised Machine Learning

- Beberapa contoh *classifier* yang sering digunakan:
  - k-Nearest Neighbors
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines (SVM)
  - …

| #2 Metode Text Classification

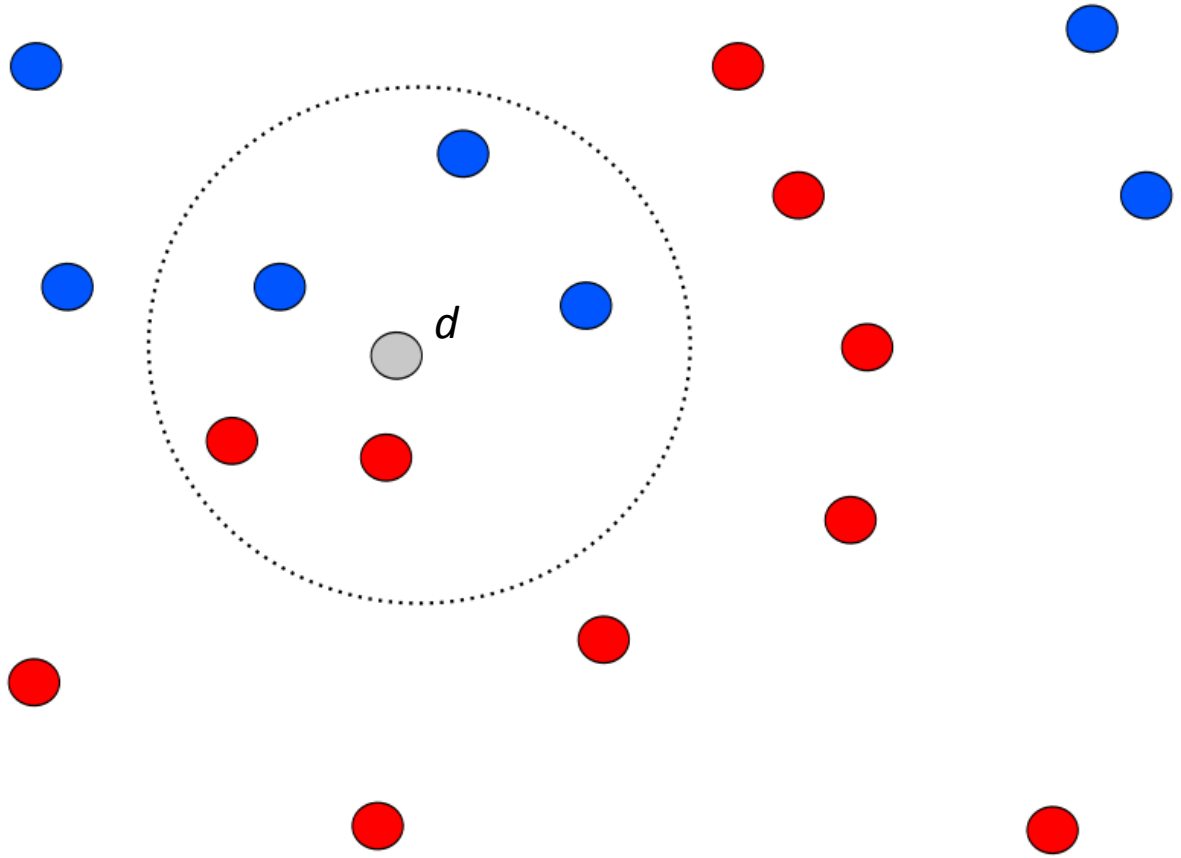






| #2 Metode Text Classification

k nearest-neighbours



- Tujuan: Mencari prediksi dari kelas untuk obyek *d*
  → yaitu kelas yang paling sering muncul di "sekitaran" *d*
- Cosine distance: sebuah *metric* yang mungkin digunakan untuk merepresentasikan "kedekatan" dari dokumen-dokumen

| #2 Metode Text Classification

k nearest-neighbours

- **Kelebihan**
  - Efektif untuk klasifikasi teks
  - Menangani kasus multi-class secara langsung
  - Tidak “membutuhkan model” (hanya menghitung representasi vektor)
- **Kekurangan**
  - Membutuhkan *fine-tuning* untuk nilai k (≈ 40 untuk klasifikasi teks)
  - Membutuhkan proses *adjustment* ketika jumlah kelas yang ada pada data tidak berimbang

| #2 Metode Text Classification

K Nearest Neighbor for Text

Training:

For each each training example <x, $c(x)$> $\in$ D

Compute the corresponding TF-IDF vector, d_x, for document x

Testing (instance y):

Compute TF-IDF vector d for document y

For each <x, $c(x)$> $\in$ D

Let $s_x = 1 - cosSim(d, d_x)$

Sort examples, x, in D by increasing value of s_x (small to big)

Let N be the first k examples in D. (*get most similar neighbors*)

Return the majority class of examples in N

| #3 Studi Kasus Text Classification - KNN

d1
Sekarang saya sedang suka memasak. Masakan kesukaan saya sekarang adalah nasi goreng. Cara memasak nasi goreng adalah nasi digoreng
Kelas A

d2
Ukuran nasi sangatlah kecil, namun saya selalu makan nasi
Kelas A

d3
Nasi berasal dari beras yang ditanam di sawah. Sawah berukuran kecil hanya bisa ditanami sedikit beras
Kelas B

d4
Mobil dan bus dapat mengangkut banyak penumpang. Namun, bus berukuran jauh lebih besar dari mobil, apalagi mobil-mobilan
Kelas B

d5
Bus pada umumnya berukuran besar dan berpenumpang banyak, sehingga bus tidak bisa melewati persawahan
Kelas C

Dokumen baru
Nasi Goreng
Kelas = ???

Nearest neighbor

$n=3$

| #3 Studi Kasus Text Classification - KNN

TERM	D1	D2	D3	D4	D5	D_{new}
suka	0,51932239	0	0	0	0	0
masak	0,58961142	0	0	0	0	0
nasi	0,18713901	0,36624274	0,158012148	0	0	0,302697098
goreng	0,58961142	0	0	0	0	0,95308681
ukur	0	0,28150215	0	0,148958474	0,278625389	0
makan	0	0,88691758	0	0	0	0
beras	0	0	0,647708116	0	0	0
tanam	0	0	0,647708116	0	0	0
sawah	0	0	0,368755414	0	0,499782802	0
mobil	0	0	0	0,751874826	0	0
bus	0	0	0	0,347626715	0,650232416	0
angkut	0	0	0	0,469317522	0	0
tumpang	0	0	0	0,267193467	0,499782802	0

| #3 Studi Kasus Text Classification - KNN

- Hitung jarak dari dokumen baru ke masing-masing dokumen.
- Misal menghitung dokumen baru ke D1

D1	0,519322	0,589611	0,187139	0,589611	0	0	0	0	0	0	0	0	0
D_{new}	0	0	0,302697	0,953087	0	0	0	0	0	0	0	0	0

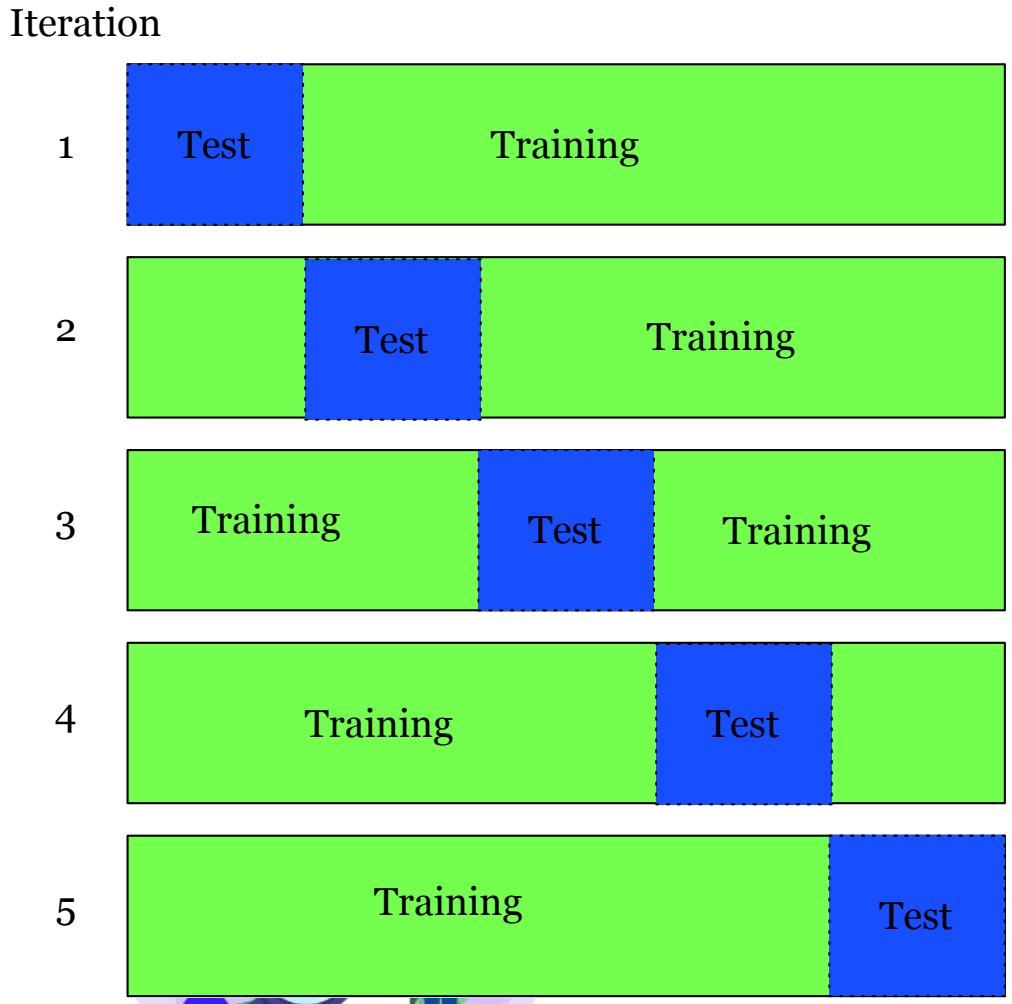
→ $d(\text{D1}, D_{new}) = 1 - cosine(\text{D1}, D_{new}) = 1 - 0{,}618597 = 0{,}381403$

Dokumen	Label	Jarak dokumen baru ke dokumen ke-n
D1	Kelas A	0,381403
D2	Kelas A	0,889139
D3	Kelas B	0,95217
D4	Kelas B	1
D5	Kelas C	1

- ✓ Urutan kelas berdasarkan nilai jarak D_n ke D_{new}, diurutkan dari kecil ke besar, adalah: {A,A,B,B,C}
- ✓ Karena N=3, maka himpunan kelas yang dipilih adaha {A,A,B}
- ✓ Kelas untuk D_{new} adalah A (kelas yang paling banyak muncul di himpunan kelas yang terpilih)

#4 Evaluasi

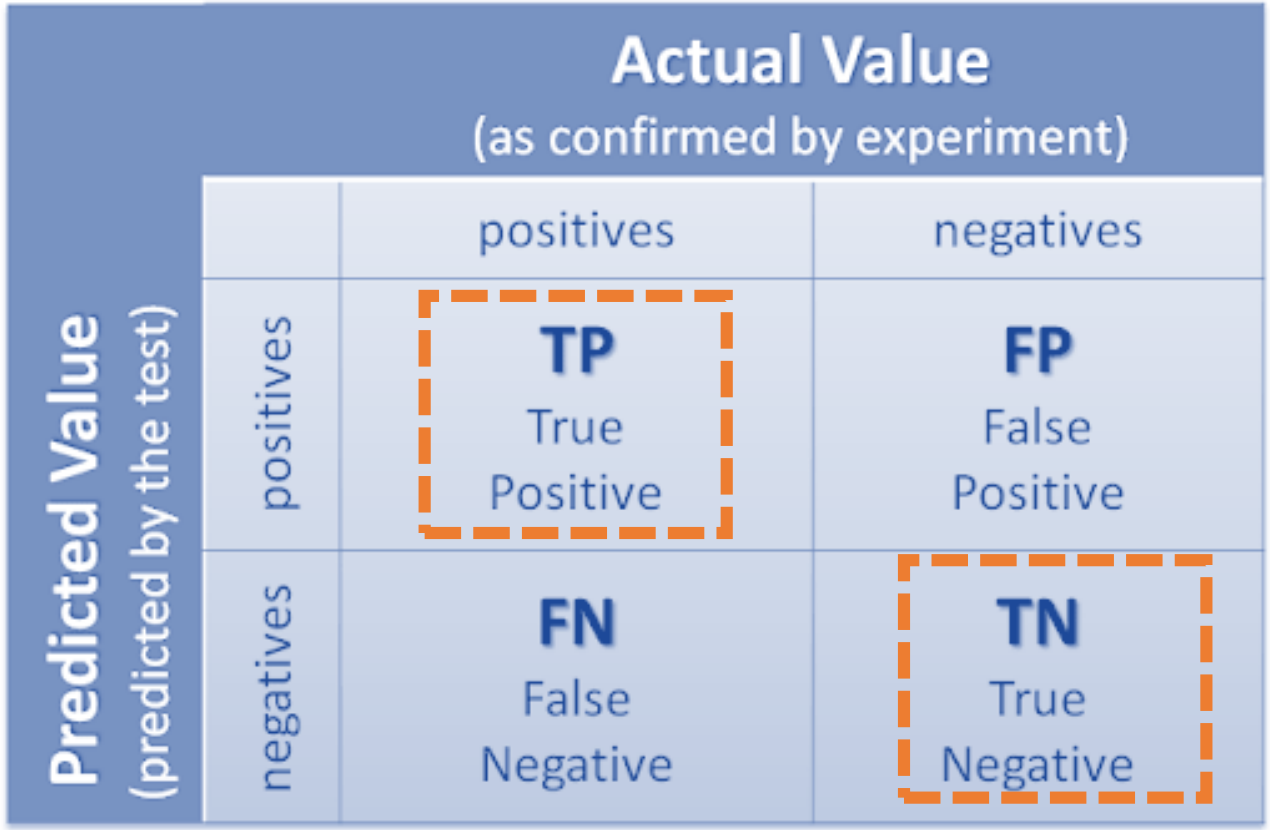
- Cross Validation



Dari: Jurafsky & Manning “Intro NLP course”, Stanford Univ

| #4 Evaluasi

- Contingency Table



- Contoh kasus : Mahasiswa DO

Nim	Status Sebenarnya	Hasil Prediksi
001	Tidak DO	Tidak DO
002	Tidak DO	Tidak DO
003	Tidak DO	Tidak DO
004	Tidak DO	DO
005	Tidak DO	DO
006	DO	Tidak DO
007	DO	DO
008	DO	DO
009	DO	DO
010	DO	DO

•**True Positive (TP)** : kasus dimana mahasiswa diprediksi (Positif) DO, memang benar(True) DO. Jadi nilai TP = 4
•**True Negative (TN)** : kasus dimana mahasiswa diprediksi tidak(Negatif) DO dan sebenarnya mahasiswa tersebut memang (True) tidak DO. Jadi TN = 3
•**False Positve (FP)** : kasus dimana mahasiswa yang diprediksi positif DO, ternyata tidak DO. Prediksinya salah (False). Nilai FP = 2
•**False Negatif (FN)**: kasus dimana mahasiswa yang diprediksi tidak DO (Negatif), tetapi ternyata sebenarnya(TRUE) DO. Jadi FN = 1

| #4 Evaluasi

Accuracy

Merupakan rasio prediksi Benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan "Berapa persen mahasiswa yang benar diprediksi DO dan Tidak DO dari kesuluruhan mahasiswa"

Akurasi = (TP + TN) / (TP+FP+FN+TN)

pada contoh kasus di atas, Akurasi = (4+3) / (4+2+1+3) = 7/10 = 70%

Precission

Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. Precission menjawab pertanyaan "Berapa persen mahasiswa yang benar DO dari keseluruhan mahasiswa yang diprediksi DO?"

Precission = (TP) / (TP+FP)

pada contoh kasus di atas, Precission = 4 / (4+2) = 4/6 = 67%.

Recall (Sensitifitas)

Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Recall menjawab pertanyaan "Berapa persen mahasiswa yang diprediksi DO dibandingkan keseluruhan mahasiswa yang sebenarnya DO".

Recall = (TP) / (TP + FN)

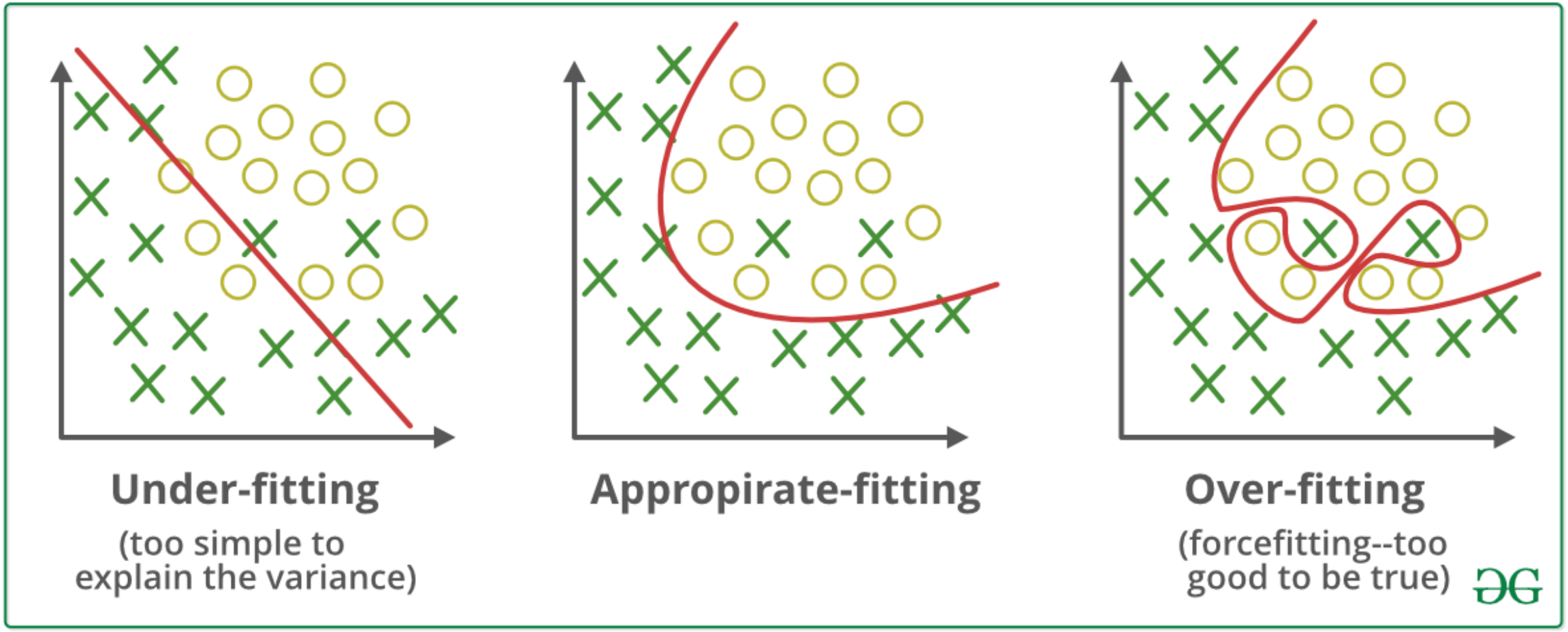
pada contoh kasus di atas Recall = 4/(4+1) = 4/5 =80%.

Bahan bacaan :

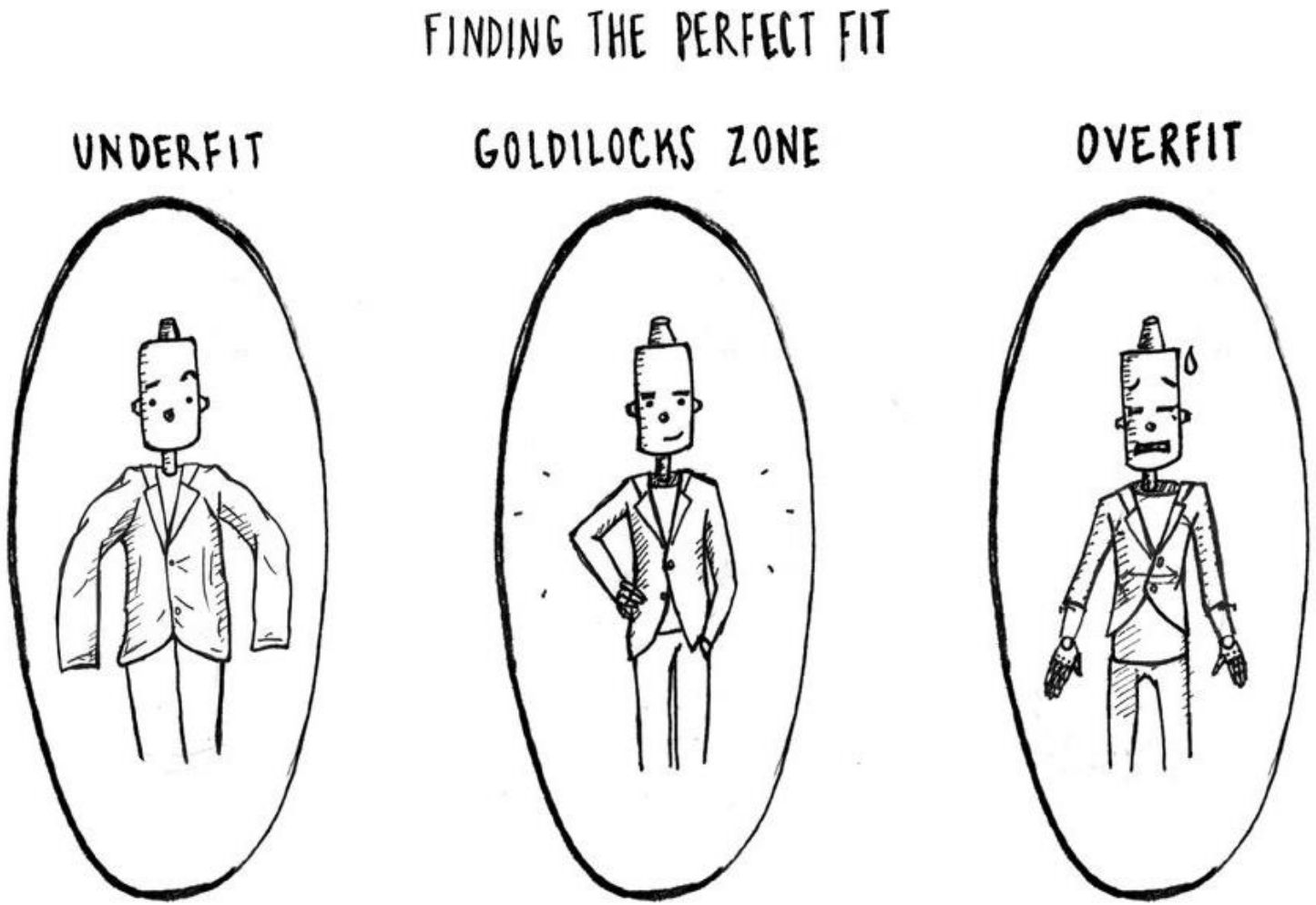
https://medium.com/@rey1024/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

| #4 Evaluasi

Underfitting & Overfitting



| #4 Evaluasi



Underfitting
& **Overfitting**

Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR