

Algoritma Naïve Bayes untuk Klasifikasi Dokumen Teks

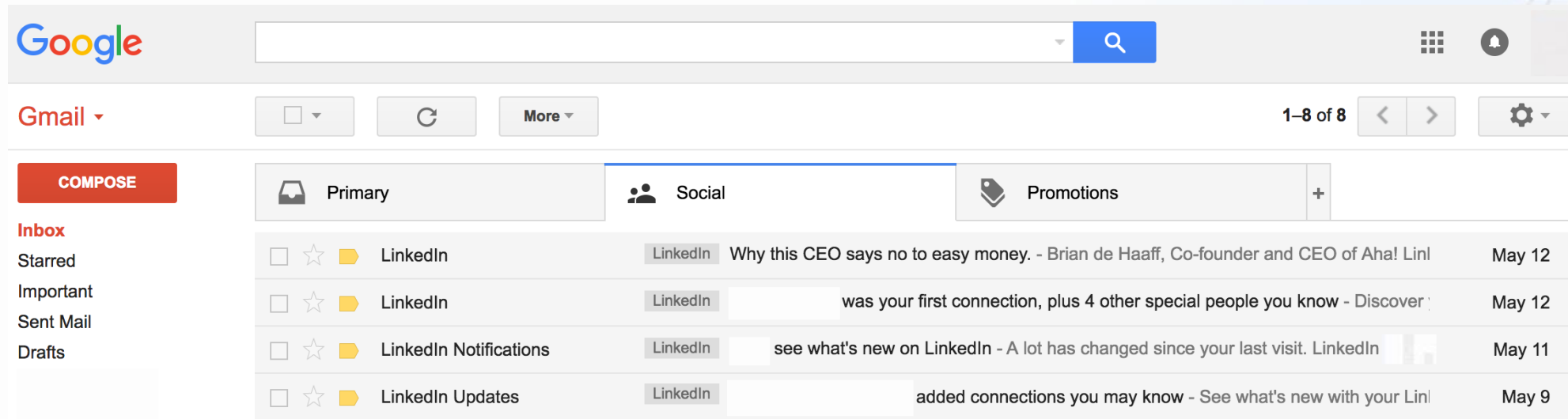
Kecerdasan Artifisial(CIF63310 / 2 sks)

Outline



- Permasalahan Klasifikasi
- Naïve Bayes
- Multinomial Naïve Bayes
- Evaluasi
- Permasalahan Umum Klasifikasi

Pengelompokan Email



Google

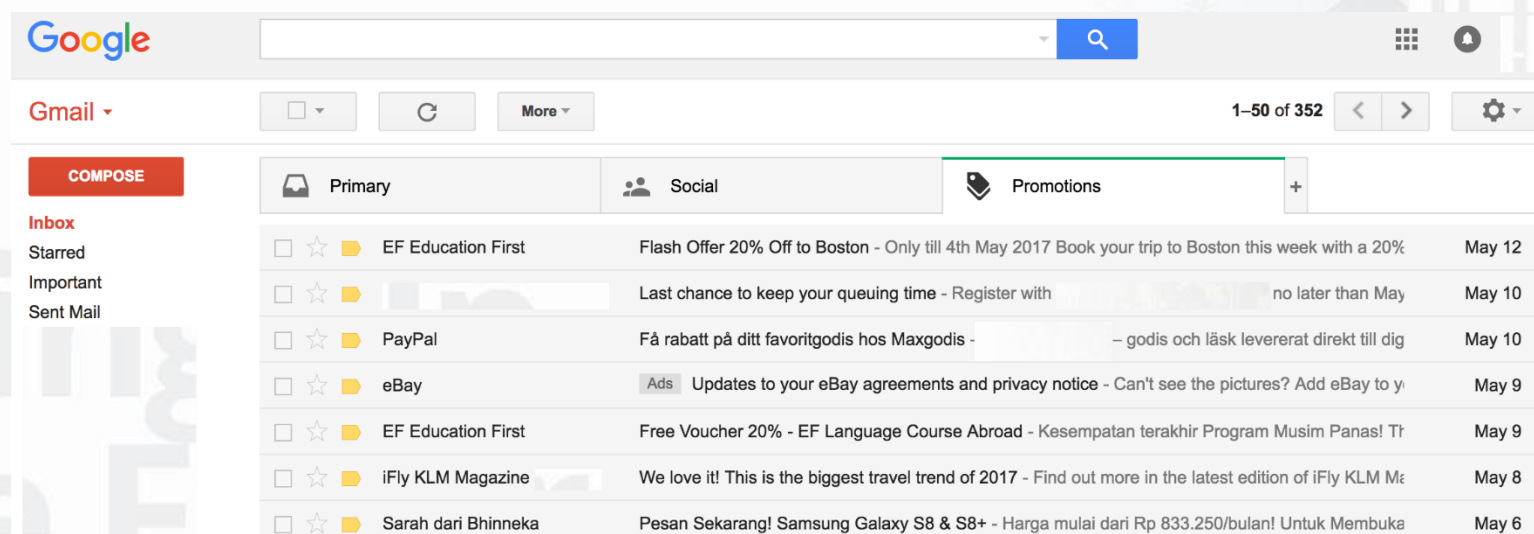
Gmail ▾

COMPOSE

Inbox
Starred
Important
Sent Mail
Drafts

Primary Social Promotions +

<input type="checkbox"/>	★	LinkedIn	LinkedIn	Why this CEO says no to easy money. - Brian de Haaff, Co-founder and CEO of Aha! Linl	May 12
<input type="checkbox"/>	★	LinkedIn	LinkedIn	was your first connection, plus 4 other special people you know - Discover	May 12
<input type="checkbox"/>	★	LinkedIn Notifications	LinkedIn	see what's new on LinkedIn - A lot has changed since your last visit. LinkedIn	May 11
<input type="checkbox"/>	★	LinkedIn Updates	LinkedIn	added connections you may know - See what's new with your Linl	May 9



Google

Gmail ▾

COMPOSE

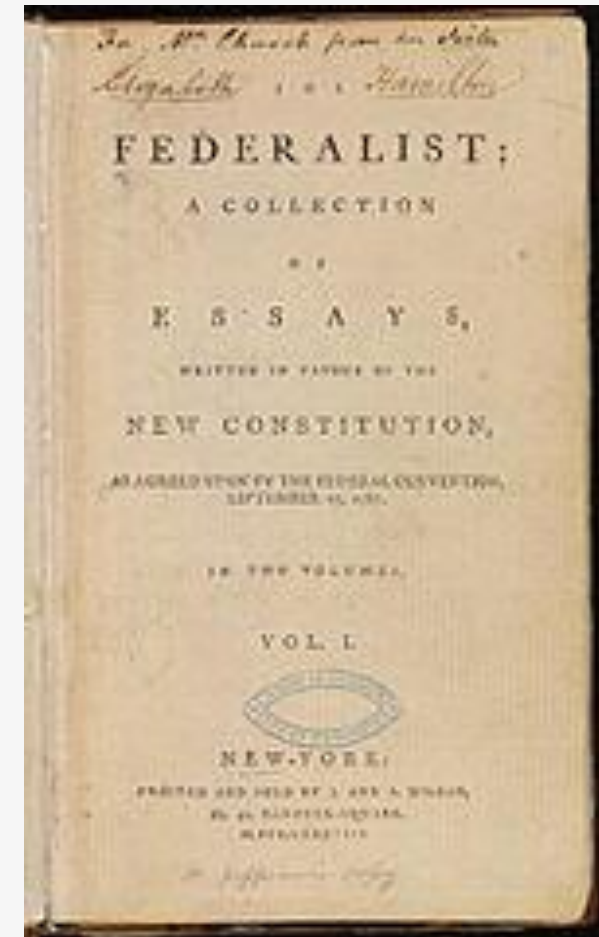
Inbox
Starred
Important
Sent Mail

Primary Social Promotions +

<input type="checkbox"/>	★	EF Education First		Flash Offer 20% Off to Boston - Only till 4th May 2017 Book your trip to Boston this week with a 20%	May 12
<input type="checkbox"/>	★			Last chance to keep your queuing time - Register with	May 10
<input type="checkbox"/>	★	PayPal		Få rabatt på ditt favoritgodis hos Maxgodis -	May 10
<input type="checkbox"/>	★	eBay	Ads	Updates to your eBay agreements and privacy notice - Can't see the pictures? Add eBay to y	May 9
<input type="checkbox"/>	★	EF Education First		Free Voucher 20% - EF Language Course Abroad - Kesempatan terakhir Program Musim Panas! Tr	May 9
<input type="checkbox"/>	★	iFly KLM Magazine		We love it! This is the biggest travel trend of 2017 - Find out more in the latest edition of iFly KLM M	May 8
<input type="checkbox"/>	★	Sarah dari Bhinneka		Pesan Sekarang! Samsung Galaxy S8 & S8+ - Harga mulai dari Rp 833.250/bulan! Untuk Membuka	May 6

Deteksi Penulis

- Pada tahun 1787-1788 terbit kumpulan esai berjudul **The Federalist Papers**, berisi ajakan untuk meratifikasi konstitusi AS
- John Jay menulis 5 paper, Alexander Hamilton menulis 51, dan James Madison 14.
- Namun, ada 15 paper yang tidak diketahui siapa penulisnya (Hamilton atau Madison)
- Tetapi ditulis secara **Anonim** oleh **Publius**
- Pada tahun 1963, para peneliti Frederick Mosteller dan David Wallace menggunakan metode statistik Bayesian untuk menentukan siapa penulis sebenarnya dari 15 esai yang tidak diketahui.
- Hasil analisis mereka menunjukkan bukti kuat bahwa James Madison adalah penulis dari esai-esai tersebut.



Pengkategorian Berita



OLAHRAGA

EKONOMI

OTOMOTIF

WISATA

Analisis Sentimen



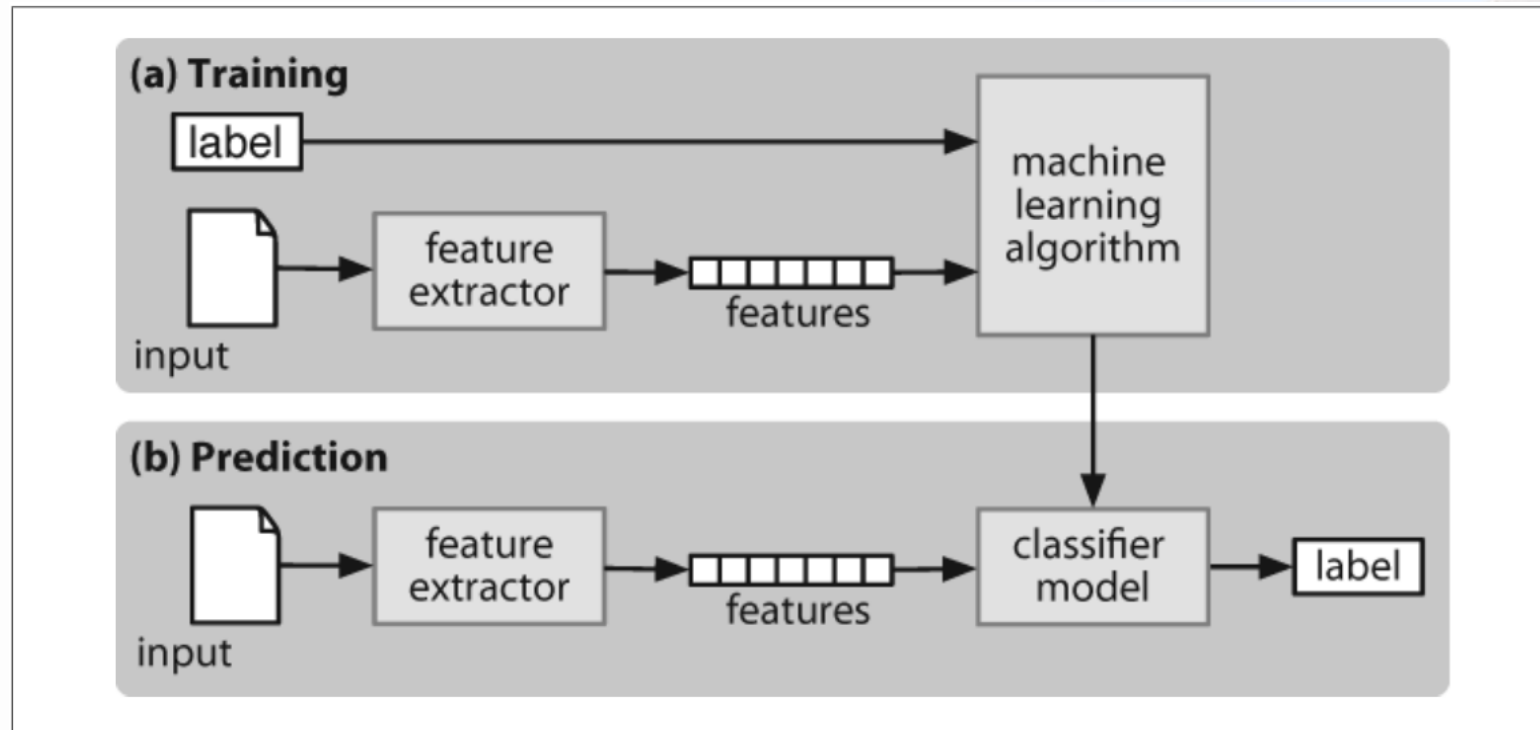
Untuk sebuah laptop dengan berat hanya 1.2kg, desain sangat bagus, build quality juga bisa dibilang wow, spesifikasi cukup untuk kegiatan harian, ini akan sangat pas dibeli sebagai laptop hadiah bagi anak sekolah atau barangkali pasangan yang sedang mengerjakan skripsi, istri yang aktif jualan online, dan anda yang betah nonton drama korea sampai berjam-jam sepanjang malam. Daya tahan baterai dan kualitas layar serta audio-nya lumayan bagus.



Definisi

- Klasifikasi merupakan pemilihan **label/kategori** yang tepat untuk suatu input
- Label/kategori umumnya sudah ditentukan di awal
- Input:
 - Dokumen d
 - Sekumpulan kategori $C = \{c_1, c_2, \dots, c_j\}$
 - Data latih sebanyak m dokumen yang telah diberi label secara manual $(d_1, c_1), \dots, (d_m, c_m)$
- Output:
 - *Classifier* yang telah dilatih $\gamma: d \rightarrow c$
- Klasifikasi merupakan bagian dari *supervised learning*

Diagram Alir Klasifikasi



Algoritma Klasifikasi

- Berbagai algoritma klasifikasi dapat digunakan untuk melakukan klasifikasi dokumen:
 - Naïve Bayes
 - Logistic Regression
 - Support Vector Machine (SVM)
 - Maximum Entropy
 - Decision Tree
 - K-Nearest Neighbors (KNN)
 - dll

Naïve Bayes

Naïve Bayes

- Metode klasifikasi sederhana menggunakan Teorema Bayes
- Fitur yang digunakan: bag of words

Representasi Bag of Words



Untuk sebuah laptop dengan berat hanya 1.2kg, desain sangat **bagus**, build quality juga bisa dibilang **wow**, spesifikasi **cukup** untuk kegiatan harian, ini akan sangat **pas** dibeli sebagai laptop hadiah bagi anak sekolah atau barangkali pasangan yang sedang mengerjakan skripsi, istri yang aktif jualan online, dan anda yang betah nonton drama korea sampai berjam-jam sepanjang malam. Daya tahan baterai dan kualitas layar serta audio-nya lumayan **bagus**.

Representasi Bag of Words

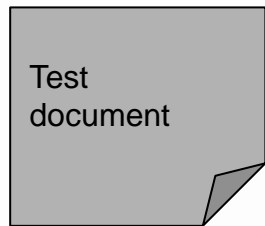
Untuk sebuah laptop dengan berat hanya 1.2kg, desain sangat **bagus**, build quality juga bisa dibilang **wow**, spesifikasi **cukup** untuk kegiatan harian, ini akan sangat **pas** dibeli sebagai laptop hadiah bagi anak sekolah atau barangkali pasangan yang sedang mengerjakan skripsi, istri yang aktif jualan online, dan anda yang betah nonton drama korea sampai berjam-jam sepanjang malam. Daya tahan baterai dan kualitas layar serta audio-nya lumayan **bagus**.

Kata	Frekuensi
bagus	2
wow	1
cukup	1
pas	1



Klasifikasi Menggunakan Bag of Words

?



Test
document

parser
language
label
translation
...

Machine
Learning

learning
training
algorithm
shrinkage
network...

NLP

parser
tag
training
translation
language...

Garbage
Collection

garbage
collection
memory
optimization
region...

Planning

planning
temporal
reasoning
plan
language...

Naïve Bayes pada Dokumen

- Untuk dokumen d dan kategori c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes pada Dokumen

- Pemilihan kelas yang paling sesuai

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c|d)$$

MAP adalah “maximum a posteriori” = kelas yang paling mungkin

$$= \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

Teorema Bayes

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Penyebut
dibuang

Naïve Bayes pada Dokumen

- Pemilihan kelas dari dokumen dengan n jumlah fitur

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d \mid c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n \mid c)P(c)$$

Dokumen
dengan fitur
 $x_1 \dots x_n$

Penentuan kelas hanya dapat dilakukan jika terdapat data latih yang berukuran sangat besar

Multinomial Naïve Bayes

Multinomial Naïve Bayes

- **Asumsi fitur:** Fitur menggunakan bag of words, posisi kata tidak diperhatikan
- **Conditional Independence:** Diasumsikan fitur-fitur yang ada bersifat independen

$$P(x_1, \dots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \dots \bullet P(x_n \mid c)$$

Multinomial Naïve Bayes

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n \mid c) P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{x \in X} P(x \mid c)$$

Kelas yang paling mungkin, dihitung menggunakan multinomial naïve bayes

Maximum Likelihood

- Langkah pertama: estimasi maximum likelihood
 - Gunakan frekuensi pada data (dokumen)

$$\hat{P}(c_j) = \frac{\text{doc_count}(C = c_j)}{N_{doc}}$$

Jumlah dokumen dengan kelas c_j

Total jumlah dokumen

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

Jumlah kata w_i dengan kelas c_j

Jumlah kata pada kelas c_j

Permasalahan pada Maximum Likelihood

- Estimasi maximum likelihood dapat bernilai 0 jika dokumen latih tidak memiliki suatu kata pada dokumen uji
- Misal terdapat kata “jelek” pada data uji, dan tidak ada kata “jelek” pada data latih dengan kelas “positif”
- Nilai peluang suatu kelas akan bernilai 0 pula

$$\hat{P}(\text{"jelek"}|\text{positif}) = \frac{\text{count}(\text{"jelek"}, \text{positif})}{\sum_{w \in V} \text{count}(w, \text{positif})} = 0$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{x \in X} P(x | c)$$

Laplace Smoothing

- Untuk menghindari nilai 0 pada maximum likelihood, gunakan Laplace Smoothing
- $count(w_i, c_j)$ = Jumlah kemunculan kata w_i dalam dokumen yang berlabel c_j
- $\sum_{w \in V} count(w, c_j)$ = Total jumlah kata dalam semua dokumen berlabel c_j
- $|V|$: Ukuran kosakata, yaitu jumlah semua kata unik dalam data pelatihan.

$$\hat{P}(w_i | c_j) = \frac{count(w_i, c_j) + 1}{(\sum_{w \in V} count(w, c_j)) + |V|}$$

Training Multinomial Naïve Bayes

Tahapan training pada Multinomial Naïve Bayes

1. Dari data latih, ekstrak *Vocabulary*, yaitu kata-kata yang relevan dengan kelas yang ada

2. Hitung $P(c_j)$

Untuk setiap kelas c_j yang ada, lakukan

- $docs_j$ = semua dokumen dengan kelas c_j

- $$P(c_j) = \frac{\text{jumlah } docs_j}{\text{jumlah semua dokumen}}$$

- *Contoh: Jika terdapat 100 dokumen dalam data pelatihan:*

- 30 dokumen di kelas A $\rightarrow P(A)=30/100=0.3$
- 70 dokumen di kelas B $\rightarrow P(B)=70/100=0.7$

Training Multinomial Naïve Bayes

3. Hitung $P(w_k | c_j)$

- $Text_j$ = gabungan dari semua dokumen $docs_j$
- Untuk setiap kata w_k pada *Vocabulary*
 - n_k = jumlah kemunculan w_k pada $Text_j$

$$P(w_k | c_j) \propto \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

α merupakan parameter smoothing
Biasanya bernilai 1, tapi bisa bernilai lain

Training Multinomial Naïve Bayes

3. Hitung $P(w_k | c_j)$

Misalkan:

$$P(w_k | c_j) = \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

- w_k = "machine"
- $n_k = 3$ (kata "machine" muncul 3 kali di *Textj*)
- $n = 50$ (total jumlah kata dalam *Textj*)
- $|Vocabulary| = 20$ (jumlah total kata unik)
- $\alpha = 1$

$$P(w_k | c_j) = \frac{3 + 1}{50 + 1 * 20} = \frac{4}{70} \approx 0.057$$

Training Multinomial Naïve Bayes

3. Tambahkan satu kata, yaitu kata “unknown” w_u

$$\hat{P}(w_u | c) = \frac{\text{count}(w_u, c) + 1}{\sum_{w \in V} \text{count}(w, c) + |V| + 1}$$

$$= \frac{1}{\sum_{w \in V} \text{count}(w, c) + |V| + 1}$$

Naïve Bayes dan Model Bahasa



- Naïve bayes dapat menggunakan fitur apapun
 - URL, alamat email, kamus
- Pada contoh di slide ini, fitur yang digunakan adalah kata dalam dokumen
- Oleh karena itu, Naïve bayes memiliki kemiripan dengan model bahasa

Naïve Bayes dan Model Bahasa

- Setiap kelas merupakan model bahasa unigram
- Menghitung peluang setiap kata: $P(word|c)$
- Menghitung peluang setiap kalimat: $P(s|c) = \prod P(word|c)$

Kelas *positif*

0.1	I	<u>I</u>	<u>love</u>	<u>this</u>	<u>fun</u>	<u>film</u>
0.1	love	0.1	0.1	.05	0.01	0.1
0.01	this					
0.05	fun					
0.1	film					
...						

$$P(s|\text{positif}) = P(I|\text{positif}) \cdot P(\text{love}|\text{positif}) \cdot P(\text{this}|\text{positif}) \cdot P(\text{fun}|\text{positif}) \cdot P(\text{film}|\text{positif}) = 0.0000005$$

Naïve Bayes dan Model Bahasa

- Kelas manakah yang paling tepat untuk sebuah kalimat?

Model positif

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model negatif

0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1 0.2	0.1 0.001	0.01 0.01	0.05 0.005	0.1 0.1

$$P(s|\text{positif}) > P(s|\text{negatif})$$

Contoh Perhitungan Naïve Bayes

$$\hat{P}(c_j) = \frac{\text{doc_count}(C = c_j)}{N_{\text{doc}}}$$

Prior

s:
 $P(c) = \frac{3}{4} \frac{1}{4}$

$$\hat{P}(w_i|c_j) = \frac{\text{count}(w_i, c_j) + 1}{(\sum_{w \in V} \text{count}(w, c_j)) + |V|}$$

	Do c	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Peluang bersyarat:

$$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$$

$$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$$

Menentukan kelas:

$$P(c|d5) \propto \frac{3}{4} * (\frac{3}{7})^3 * \frac{1}{14} * \frac{1}{14} \\ \approx 0.0003$$

$$P(j|d5) \propto \frac{1}{4} * (\frac{2}{9})^3 * \frac{2}{9} * \frac{2}{9} \\ \approx 0.0001$$

Contoh Perhitungan Naïve Bayes

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

A. Hitung Prior Probability ($P(c)$)

$$\hat{P}(c_j) = \frac{\text{doc_count}(C = c_j)}{N_{\text{doc}}}$$

Priors:

$$P(c) = \frac{3}{4}$$
$$P(j) = \frac{1}{4}$$

B. Hitung Likelihood ($P(w|c)$)

$$\hat{P}(w_i|c_j) = \frac{\text{count}(w_i, c_j) + 1}{(\sum_{w \in V} \text{count}(w, c_j)) + |V|}$$

$$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$$

$$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$$

C. Hitung Probabilitas untuk Dokumen Uji

$$P(c|d5) = P(C) * P(\text{Chinese}|c) * P(\text{Tokyo}|c) * P(\text{Japan}|c)$$

$$P(c|d5) = \frac{3}{4} * \left(\frac{3}{7}\right)^3 * \frac{1}{14} * \frac{1}{14} \approx 0.0003$$

$$P(j|d5) = P(j) * P(\text{Chinese}|j) * P(\text{Tokyo}|j) * P(\text{Japan}|j)$$

$$P(j|d5) = \frac{1}{4} * \left(\frac{2}{9}\right)^3 * \frac{2}{9} * \frac{2}{9} \approx 0.0001$$

Evaluasi

Confusion matrix:

Binary classification

	Ground Truth Benar	Ground Truth Salah
Prediksi Benar/Terpilih	<i>tp</i>	<i>fp</i>
Prediksi Salah/Tidak terpilih	<i>fn</i>	<i>tn</i>

- ***Keterangan:***

- *tp*: True Positive
- *fn*: False Negative
- *fp*: False Positive
- *tn*: True Negative
- Hasil prediksi didapatkan dari sistem
- Hasil *ground truth/true value* didapatkan dari pakar

Precision and Recall

- **Precision/Positive Predictive Value:**
 - % data bernilai benar dari data yang terpilih/diprediksi
- **Recall/Sensitivity/Hit Rate/True Positive Rate (TPR):**
 - % data diprediksi benar dari seluruh data yang benar (termasuk yang tidak terpilih)
- **Specifity/True Negative Rate (TNR):**
 - % dari data salah yang benar diprediksi
- Menghitung akurasi (*precision*) saja tidak cukup!!!

Precision and Recall



	Ground Truth Benar	Ground Truth Salah
Prediksi Benar/Terpilih	tp	fp
Prediksi Salah/Tidak terpilih	fn	tn

$$Precision = \frac{tp}{tp + fp}$$

$$Recall/Sensitivity = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{tn + fp}$$

F-measure

- Jika menginginkan keseimbangan antara P/R, gunakan balanced F-measure

$$F = \frac{2PR}{P+R}$$

DISKUSI

Building Up
Noble Future

TERIMA KASIH