TEXT MINING

Klasifikasi Dokumen dengan Naïve Bayes

Team Teaching

FILKOM
FAKULTAS ILMU KOMPUTER
Fakultas Ilmu Komputer Universitas Brawijaya

Klasifkasi Teks dengan Naïve Bayes

- Merupakan metode klasifikasi yang sederhana (“naïve”) berdasarkan aturan Bayes
- Umumnya menggunakan Bag of Words sebagai fitur

Klasifkasi Teks dengan Naïve Bayes

- Kategori dari dokumen uji *d* adalah kategori yang memiliki nilai probabilitas posterior *P(c|d)* terbesar

$$C(d) = arg \max_{c} P(c|d)$$

- Untuk dokumen *d* dan kategori *c, probabilitas posterior* bisa dihitung*:*

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

- *P(c|d)* disebut **posterior** atau peluang kelas *c* diberikan dokumen *d*
- *P(c)* disebut **prior,** atau peluang awal munculnya kategori *c*
- *P(d|c)* disebut ***Likelihood*** atau ***conditional probability***
- *P(d)* disebut ***evidence*** atau peluang munculnya dokumen *d*

- Masing-masing kategori akan dihitung nilai posteriornya
- Jika kita memiliki 3 kategori {*c1, c2,* dan *c3*}, maka akan dihitung
- $P(c1|d) = \frac{P(c1)*P(d|c1)}{P(d)}$
- $P(c2|d) = \frac{P(c2)*P(d|c2)}{P(d)}$
- $P(c3|d) = \frac{P(c3)*P(d|c3)}{P(d)}$
- Selanjutnya, akan dipilih kategori yang memiliki nilai posterior terbesar sebagai kategori dari dokumen *d*

- Misal $P(c1|d) = 0.03$, $P(c2|d) = 0.01$ dan $P(c3|d) = 0.006$, maka dokumen *d* akan masuk pada kategori *c1*
- Meskipun probabilitas posteriornya kecil, kategori *c1* tetap dipilih karena nilainya paling besar dibandingkan probabilitas posterior kategori lain
- Kita hanya melihat mana kategori dengan peluang tersbesar tanpa peduli besar nilainya
- Oleh karena itu, *P(d)* bisa dihilangkan karena *P(d)* selalu bernilai sama untuk semua kategori

- Contoh:
- $P(c1|d) = \frac{P(c1)*P(d|c1)}{P(d)} = \frac{0.006}{P(d)}$
- $P(c2|d) = \frac{P(c2)*P(d|c2)}{P(d)} = \frac{0.002}{P(d)}$
- $P(c3|d) = \frac{P(c3)*P(d|c3)}{P(d)} = \frac{0.0012}{P(d)}$
- Karena nilai *P(d)* selalu sama untuk semua kategori, tanpa memperhitungkan *P(d)* kita sudah tahu bahwa *P(c1|d)* akan mempunyai nilai probabilitas terbesar.
- Oleh karena itu, kita sudah bisa memastikan bahwa dokumen *d* akan masuk kategori *c1*

- Bukti: Misal P(d)=0.2, kita hitung posteriornya:
- $P(c1|d) = \frac{P(c1)*P(d|c1)}{P(d)} = \frac{0.006}{0.2} = 0.03$
- $P(c2|d) = \frac{P(c2)*P(d|c2)}{P(d)} = \frac{0.002}{0.2} = 0.01$
- $P(c3|d) = \frac{P(c3)*P(d|c3)}{P(d)} = \frac{0.0012}{P(d)} = 0.006$
- Terbukti, *P(c1|d)* memang mempunyai nilai probabilitas terbesar.
- Oleh karena itu, kita tidak perlu melibatkan P(d) ke dalam perhitungan untuk penentuan kategori

- Oleh karena itu, Formula Naïve Bayes untuk klasifikasi cukup:

$$P(c|d) = P(c) * P(d|c)$$

- *P(c|d)* disebut **posterior** atau peluang kelas *c* diberikan dokumen *d*
- *P(c)* disebut **prior,** atau peluang awal munculnya kategori *c*
- *P(d|c)* disebut ***Likelihood*** atau ***conditional probability***

Prior probability

- Sekarang, mari fokus pada *Prior*:

$$P(c|d) = P(c) * P(d|c)$$

- Prior adalah probablitas awal dari kelas *c*
- Perhitungan Prior hanya berdasarkan pada data latih

$$P(c) = \frac{Nc}{N}$$

- *Nc = Jumlah dokumen pada data latih yang masuk pada kategori c*
- *N = Jumlah dokumen pada data latih*

Prior probability

- Misal, kita memiliki 10 dokumen training dan 3 kategori
- *D1, D2, D3* masuk kategori *C1*
- *D4, D5* masuk kategori *C2*
- *D6, D7, D8, D9, D10* masuk kategori *C3*

$$P(c) = \frac{Nc}{N}$$

- *Maka* $P(c1) = \frac{3}{10}$
- $P(c2) = \frac{2}{10}$ dan
- $P(c3) = \frac{5}{10}$

Conditional probability

- Sekarang, mari fokus pada *Conditional Probability*:

$$P(c|d) = P(c) * P(d|c)$$

- Dokumen *d* yang akan diklasifkasi terdiri dari beberapa kata *w*
- Probabilitas masing-masing fitur atau kata *P(w|c)* diasumsikan independen (oleh karena itu disebut "Naïve")
- Misal dokumen *d* terdiri dari 4 kata *{w1, w2, w3, w4}*
- *Conditional probability* dokumen *d* merupakan hasil perkalian *Conditional probability* masing-masing fitur:

$$P(d|c) = P(w1, w2, w3, w4|c) = P(w1|c) * P(w2|c) * P(w3|c) * P(w4|c)$$

Conditional probability

- Misal dokumen *d* terdiri dari *n* kata *{w1, w2, …, wn}*
- *Conditional probability* dokumen *d* bisa kita hitung:

$$P(d|c) = P(w1, w2, \dots, wn|c) = P(w1|c) * P(w2|c) * \cdots * P(wn|c)$$

atau

$$P(d|c) = P(w1, w2, \dots, wn|c) = \prod_{w \in d} P(w|c)$$

Conditional probability

- Oleh karena itu, Formula Naïve Bayes untuk klasifikasi dokumen menjadi:

$$P(c|d) = P(c) * \prod_{w \in d} P(w|c)$$

- *P(c|d)* disebut **posterior** atau peluang kelas *c* diberikan dokumen *d*
- *P(c)* disebut **prior,** atau peluang awal munculnya kategori *c*
- $\prod_{w \in d} P(w|c)$ adalah perkalian dari *conditional probability* masing-masing fitur atau kata yang terdapat pada dokumen *d*

Conditional probability

- Sekarang, mari fokus pada *Conditional Probability* fitur:

$$P(c|d) = P(c) * \prod_{w \in d} P(w|c)$$

- Untuk menghitung Conditional Probability, kita bisa menggunakan salah satu dari 3 model:
  - Bernoulli
  - Multinomial
  - Gaussian

Bernoulli Model

- Disebut juga *Multi-variate Bernoulli* model
- Berdasarkan pada data biner: menggunakan Binary Term Frequency

$$P(w|c) = b_t * \frac{Nc(w)}{Nc} + (1 - b_t) * (1 - \frac{Nc(w)}{Nc})$$

- Jika kata *w* ada pada dokumen berkategori *c*, $b_t = 1$
- Jika kata *w* tidak ada pada dokumen berkategori c, $b_t = 0$
- $Nc(w)$ = jumlah dokumen berkategori *c* yang mengandung kata *w*
- Nc = jumlah dokumen berkategori *c*

Bernoulli Model

- Untuk menghindari adanya probabilitas nol, biasanya dilakukan *add-one* atau *laplace smoothing.* Ada penambahan +1 pada pembilang, dan +2 pada penyebut

$$P(w|c) = b_t * \frac{Nc(w) + 1}{Nc + 2} + (1 - b_t) * (1 - \frac{Nc(w) + 1}{Nc + 2})$$

- Jika kata w ada pada dokumen berkategori c, $b_t = 1$
- Jika kata *w* tidak ada pada dokumen berkategori c, $b_t = 0$
- $Nc(w)$ = jumlah dokumen berkategori *c* yang mengandung kata w
- Nc = jumlah dokumen berkategori *c*

Multinomial Model

- Berdasarkan pada data diskrit: menggunakan Raw Term Frequency

$$P(w|c) = \frac{count(w,c)}{count(c)}$$

- $count(w,c)$ = jumlah kemunculan kata *w* pada kategori *c*
- $count(c)$ = jumlah total kemunculan semua kata pada kategori *c*

Multinomial Model

- Untuk menghindari adanya probabilitas nol, biasanya dilakukan *add-one* atau *laplace smoothing.* Ada penambahan +1 pada pembilang, dan |V| pada penyebut

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

- $count(w,c)$ = jumlah kemunculan kata *w* pada kategori *c*
- $count(c)$ = jumlah total kemunculan semua kata pada kategori *c*
- $|V|$ = jumlah term unik atau fitur

Gaussian Model

- Berdasarkan pada data kontinyu: menggunakan TF-IDF

$$P(w|c) = \frac{1}{\sqrt{2\pi{\sigma_c}^2}} e^{\left(\frac{-(w-\mu_c)^2}{2{\sigma_c}^2}\right)}$$

- π = nilai pi 3.14…
- ${\sigma_c}^2$= varians nilai fitur *w* pada kelas *c*
- w = nilai fitur *w* (nilai TF-IDF) pada data uji
- μ_c= rata-rata nilai fitur *w* pada kelas *c*

Conditional probability

- Dari ketiga model tersebut, yang paling popular dalam klasifikasi teks adalah *multinomial*
- *Multinomial model* terbukti secara umum lebih akurat dalam klasifikasi teks dibandingkan dengan model lain

Langkah-langkah Klasifikasi Teks dengan Naïve Bayes

- **Training**
  - Inputnya adalah data latih
  - Menghitung *prior* masing-masing kategori
  - Menentukan fitur berdasarkan term unik pada data latih
  - Menghitung Raw TF untuk semua fitur pada semua dokumen di data latih
  - Menghitung *conditional probability* per fitur per kategori
  - Outputnya adalah model (nilai *prior* dan *conditional probability*)
- **Testing**
  - Inputnya adalah model dan data uji
  - Menghitung *posterior* untuk masing-masing kategori dan menentukan kategori dengan nilai *posterior* terbesar
  - Outputnya adalah kategori untuk masing-masing data uji

Studi kasus klasifikasi dokumen

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 d6
Nasi Goreng Pedas
Kelas = ???

Prior probability

- *d1, d2* masuk kategori *A*
- *d3, d4* masuk kategori *B*
- *d5* masuk kategori *C*
- *Jumlah dokumen pada data latih (N) = 5*

$$P(c) = \frac{Nc}{N}$$

- *Maka* $P(A) = \frac{2}{5}$
- $P(B) = \frac{2}{5}$
- $P(C) = \frac{1}{5}$

Fase Training

Menentukan Fitur

- Term Unik sebagai Fitur. Ada 13 fitur pada kasus ini

No	TERM/FITUR
1	suka
2	masak
3	nasi
4	goreng
5	ukur
6	makan
7	beras
8	tanam
9	sawah
10	mobil
11	bus
12	angkut
13	tumpang

Studi Kasus

- Menghtung Raw TF masing-masing fitur pada masing-masing dokumen di data latih

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1

Studi Kasus

- Menghtung *conditional probability dengan Mutinomial Model* masing-masing fitur pada masing-masing kategori
- Misal, *conditional probability* fitur “ukur” pada kategori A

$$P(ukur|A) = \frac{count(ukur, A) + 1}{count(A) + |V|}$$

- $count(ukur, A)$ = jumlah kemunculan kata *ukur* pada kategori *A*
- $count(A)$ = jumlah total kemunculan semua kata pada kategori *A*
- $|V|$ = jumlah term unik atau fitur

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *A*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1

$$P(ukur|A) = \frac{count(ukur, A) + 1}{count(A) + |V|}$$

|V| = 13

count(ukur,A) = 0+1 = 1

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *A*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|A) = \frac{count(ukur, A) + 1}{count(A) + |V|}$$

|V| = 13
count(ukur,A) = 0+1 = 1
Count(A) = 11+4 = 15

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *A*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|A) = \frac{count(ukur, A) + 1}{count(A) + |V|}$$

$|V| = 13$

count(ukur,A) = 0+1 = 1

Count(A) = 11+4 = 15

$$P(ukur|A) = \frac{1+1}{15+13} = \frac{2}{28} = 0.071$$

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *B*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1

$$P(ukur|B) = \frac{count(ukur, B) + 1}{count(B) + |V|}$$

|V| = 13

count(ukur,B) = 0+1 = 1

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *B*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|B) = \frac{count(ukur, B) + 1}{count(B) + |V|}$$

|V| = 13
count(ukur,B) = 0+1 = 1
Count(B) = 7+9 = 16

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *B*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|B) = \frac{count(ukur, B) + 1}{count(B) + |V|}$$

|V| = 13
count(ukur,B) = 0+1 = 1
Count(B) = 7+9 = 16

$$P(ukur|B) = \frac{1+1}{16+13} = \frac{2}{29} = 0.069$$

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *C*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1

$$P(ukur|C) = \frac{count(ukur, C) + 1}{count(C) + |V|}$$

|V| = 13
count(ukur,C) = 1

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *C*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|C) = \frac{count(ukur, C) + 1}{count(C) + |V|}$$

|V| = 13
count(ukur,C) = 1
Count(C) = 5

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *C*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)
1	suka	2	0	0	0	0
2	masak	3	0	0	0	0
3	nasi	3	2	1	0	0
4	goreng	3	0	0	0	0
5	ukur	0	1	0	1	1
6	makan	0	1	0	0	0
7	beras	0	0	2	0	0
8	tanam	0	0	2	0	0
9	sawah	0	0	2	0	1
10	mobil	0	0	0	4	0
11	bus	0	0	0	2	2
12	angkut	0	0	0	1	0
13	tumpang	0	0	0	1	1
	Total	**11**	**4**	**7**	**9**	**5**

$$P(ukur|C) = \frac{count(ukur, C) + 1}{count(C) + |V|}$$

|V| = 13
count(ukur,C) = 1
Count(C) = 5

$$P(ukur|C) = \frac{1+1}{5+13} = \frac{2}{18} = 0.111$$

Studi Kasus

- Conditional probability fitur *ukur* pada kategori *C*

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)	P(w|A)	P(w|B)	P(w|C)
1	suka	2	0	0	0	0			
2	masak	3	0	0	0	0			
3	nasi	3	2	1	0	0			
4	goreng	3	0	0	0	0			
5	ukur	0	1	0	1	1	0.071	0.069	0.111
6	makan	0	1	0	0	0			
7	beras	0	0	2	0	0			
8	tanam	0	0	2	0	0			
9	sawah	0	0	2	0	1			
10	mobil	0	0	0	4	0			
11	bus	0	0	0	2	2			
12	angkut	0	0	0	1	0			
13	tumpang	0	0	0	1	1			
	Total	**11**	**4**	**7**	**9**	**5**			

$$P(ukur|A) = 0.071$$

$$P(ukur|B) = 0.069$$

$$P(ukur|C) = 0.111$$

Studi Kasus

- Conditional probability masing-masing fitur pada masing-masing kategori

No	TERM	D1(A)	D2(A)	D3(B)	D4(B)	D5(C)	P(w|A)	P(w|B)	P(w|C)
1	suka	2	0	0	0	0	0.107	0.034	0.056
2	masak	3	0	0	0	0	0.143	0.034	0.056
3	nasi	3	2	1	0	0	0.214	0.069	0.056
4	goreng	3	0	0	0	0	0.143	0.034	0.056
5	ukur	0	1	0	1	1	0.071	0.069	0.111
6	makan	0	1	0	0	0	0.071	0.034	0.056
7	beras	0	0	2	0	0	0.036	0.103	0.056
8	tanam	0	0	2	0	0	0.036	0.103	0.056
9	sawah	0	0	2	0	1	0.036	0.103	0.111
10	mobil	0	0	0	4	0	0.036	0.172	0.056
11	bus	0	0	0	2	2	0.036	0.103	0.167
12	angkut	0	0	0	1	0	0.036	0.069	0.056
13	tumpang	0	0	0	1	1	0.036	0.069	0.111
	Total	**11**	**4**	**7**	**9**	**5**			

Fase Testing

Studi Kasus

- Menghitung Posterior masing-masing kategori

No	TERM	P(w|A)	P(w|B)	P(w|C)
1	suka	0.107	0.034	0.056
2	masak	0.143	0.034	0.056
3	nasi	0.214	0.069	0.056
4	goreng	0.143	0.034	0.056
5	ukur	0.071	0.069	0.111
6	makan	0.071	0.034	0.056
7	beras	0.036	0.103	0.056
8	tanam	0.036	0.103	0.056
9	sawah	0.036	0.103	0.111
10	mobil	0.036	0.172	0.056
11	bus	0.036	0.103	0.167
12	angkut	0.036	0.069	0.056
13	tumpang	0.036	0.069	0.111
	Total			

Dokumen baru d6
Nasi Goreng Pedas
Kelas = ???

$$P(c|d) = P(c) * \prod_{w \in d} P(w|c)$$

$$P(c|d) = P(c) * P(nasi|c) * P(goreng|c)$$

Kata **pedas** tidak diperhitungkan karena **tidak termasuk fitur** (Tidak pernah muncul pada data latih)

Ingat-ingat lagi

- *d1, d2* masuk kategori *A*
- *d3, d4* masuk kategori *B*
- *d5* masuk kategori *C*
- *Jumlah dokumen pada data latih (N) = 5*

$$P(c) = \frac{Nc}{N}$$

- *Maka* $P(A) = \frac{2}{5} = 0.4$
- $P(B) = \frac{2}{5} = 0.4$
- $P(C) = \frac{1}{5} = 0.2$

Studi Kasus

- Menghitung Posterior masing-masing kategori

No	TERM	P(w|A)	P(w|B)	P(w|C)
1	suka	0.107	0.034	0.056
2	masak	0.143	0.034	0.056
3	nasi	0.214	0.069	0.056
4	goreng	0.143	0.034	0.056
5	ukur	0.071	0.069	0.111
6	makan	0.071	0.034	0.056
7	beras	0.036	0.103	0.056
8	tanam	0.036	0.103	0.056
9	sawah	0.036	0.103	0.111
10	mobil	0.036	0.172	0.056
11	bus	0.036	0.103	0.167
12	angkut	0.036	0.069	0.056
13	tumpang	0.036	0.069	0.111

Dokumen baru d6
Nasi Goreng Pedas
Kelas = ???

$$P(A|d6) = P(A) * P(nasi|A) * P(goreng|A)$$
$$P(A|d6) = 0.4 * 0.214 * 0.143 = 0.0122$$

$$P(B|d6) = P(B) * P(nasi|B) * P(goreng|B)$$
$$P(B|d6) = 0.4 * 0.069 * 0.034 = 0.0009$$

$$P(C|d6) = P(C) * P(nasi|C) * P(goreng|C)$$
$$P(C|d6) = 0.2 * 0.056 * 0.056 = 0.0006$$

Studi Kasus

- Menghitung Posterior masing-masing kategori

No	TERM	P(w|A)	P(w|B)	P(w|C)
1	suka	0.107	0.034	0.056
2	masak	0.143	0.034	0.056
3	nasi	0.214	0.069	0.056
4	goreng	0.143	0.034	0.056
5	ukur	0.071	0.069	0.111
6	makan	0.071	0.034	0.056
7	beras	0.036	0.103	0.056
8	tanam	0.036	0.103	0.056
9	sawah	0.036	0.103	0.111
10	mobil	0.036	0.172	0.056
11	bus	0.036	0.103	0.167
12	angkut	0.036	0.069	0.056
13	tumpang	0.036	0.069	0.111

Dokumen baru d6
Nasi Goreng Pedas
Kelas = A

Karena posterior kategori *A* **terbesar**, maka *d6* masuk ke **kategori *A***

$$P(A|d6) = P(A) * P(nasi|A) * P(goreng|A)$$
$$P(A|d6) = 0.4 * 0.214 * 0.143 = 0.0122$$

$$P(B|d6) = P(B) * P(nasi|B) * P(goreng|B)$$
$$P(B|d6) = 0.4 * 0.069 * 0.034 = 0.0009$$

$$P(C|d6) = P(C) * P(nasi|C) * P(goreng|C)$$
$$P(C|d6) = 0.2 * 0.056 * 0.056 = 0.0006$$

Latihan

Tentukan kelas dari dokumen uji

D=**Burung terbang**

Data Training:

D1: **Layang-layang terbang diangkasa (Kelas A)**

D2: **Burung- burung terbang diangkasa (Kelas B)**

D3: Banyak l**ayang-Layang** berbentuk **burung (Kelas A)**

D4: **Burung-burung di angkasa pulang** di sore hari **(Kelas B)**

D5: **Burung terbang** untuk **pulang** ke **sarang (Kelas B)**


UB
FILKOM
FAKULTAS ILMU KOMPUTER