

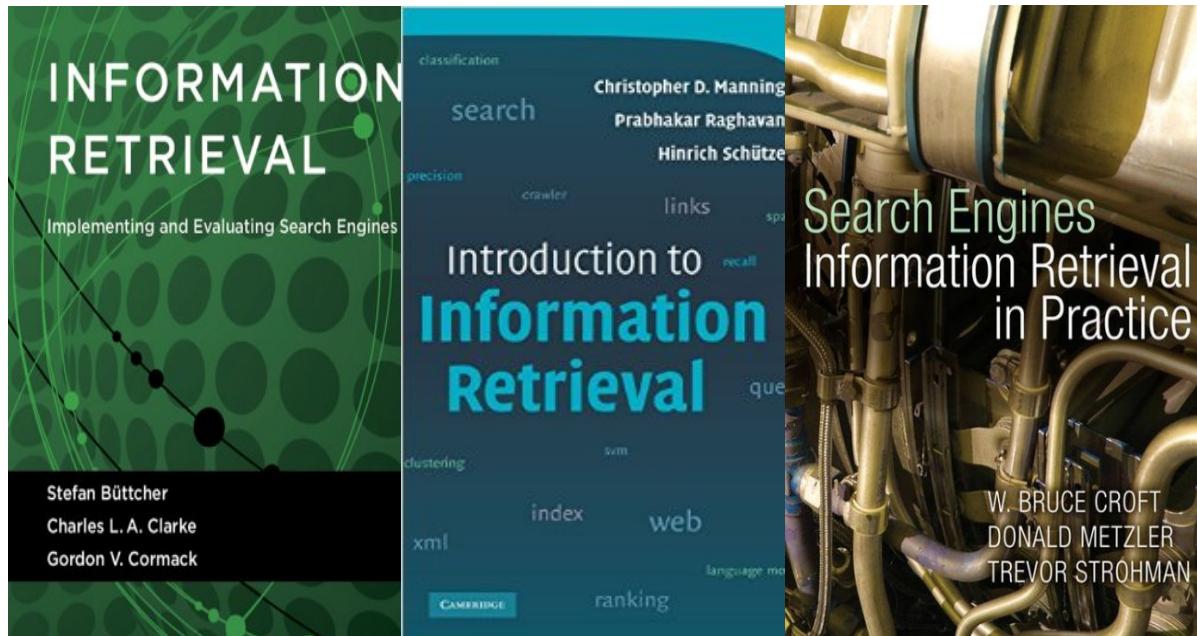
Sistem Temu Kembali Informasi

“Feature Selection”



Tim Dosen STKI

Buku Penunjang & Literatur





Penerapan Klasifikasi Teks

Google

News U.S. edition Modern

Top Stories News near you Suggested for you World U.S. Business Technology Entertainment Sports

 ESPN

Daily Word: What's next for Wisconsin?
ESPN - 1 hour ago  Each weekday, our college hoops experts discuss the biggest issues, trends and themes in college basketball. 1. What's next for Wisconsin?
In bequeathing head coach role mid-season, Bo Ryan claims final victory
SportingNews.com
Wisconsin basketball coach Bo Ryan announces retirement, effective immediately
AL.com
Opinion: Bo Ryan leaves Wisconsin: 'I'll see you down the road' Chicago Tribune

A night with the most famous bench in college basketball
USA TODAY - 8 hours ago
WASHINGTON - Monmouth made easy work out of Georgetown on Tuesday night in an 83-68 win at the Verizon Center, and the Hawks' now-famous "Bench Mob" celebrated the only way they know how: with a dance party.

No Fanfare This Time as Cavaliers Top Celtics
New York Times - 7 minutes ago
LeBron James scored 24 points to lead the Cleveland Cavaliers to an 89-77 victory over the host Boston Celtics on Tuesday night in an uneventful rematch of their bruising first-round playoff series last season.

Joey Bosa and Paxton Lynch atop McShay's first 2016 mock draft
ESPN - 1 hour ago
 College football season is nearly complete... which can mean only one thing: The NFL draft is rapidly approaching.

Related Bo Ryan » Wisconsin »



Klasifikasi Spam Mail

```
Received: from 192.168.1.100 ([65.202.85.3]) by pacific-carrier-annex.mit.edu  
        (8.9.2/8.9.2) with SMTP id AAA06179;  
        Mon, 11 Jun 2001 00:39:32 -0400 (EDT)  
From: [some forged email address]  
Message-ID: <200106110439.AAA06179@pacific-carrier-annex.mit.edu>  
Subject: I am as shocked as you!  
Date: Sun, 10 Jun 01 00:32:35 Pacific Daylight Time  
X-Priority: 3  
X-MSMailPriority: Normal  
Importance: Normal  
MIME-Version: 1.0  
Content-Type: multipart/mixed;  
        boundary="-----_NextPart_000_018C_01BD9940.715D52A0"
```

<HTML>
<BODY>

Spam=True/False

Some of the most beautiful women in the world bare it all for you.Denise Richard
s, Britney Spears, Jessica Simpson, and many more.<A HREF="http://216.130.166.1
88/index.html">CLICK HERE FOR NUDE CELEBS

</BODY></HTML >



Ketentuan

○ Fitur Ekstraksi

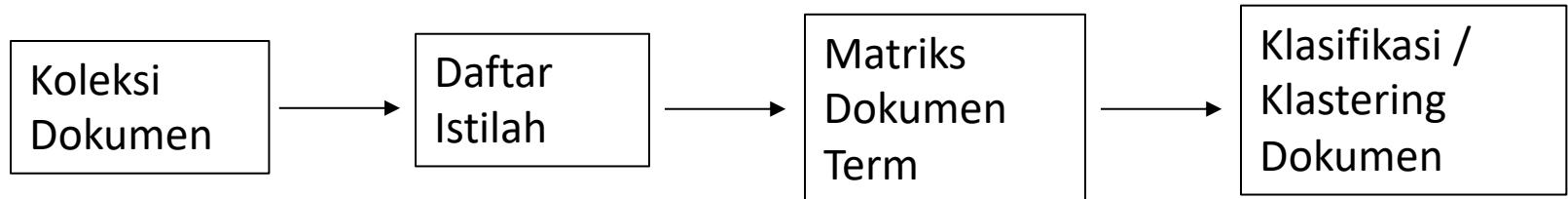
- Buat fitur dari dokumen.

○ Fitur Seleksi

- Cobalah untuk menemukan fitur subset terbaik.



Langkah Umum Klasifikasi Teks / Klastering Teks





Masalah

- Dimensi tinggi dari ruang fitur.
- Puluhan / Ratusan ribu kata.
- Tidak semua fitur membantu.

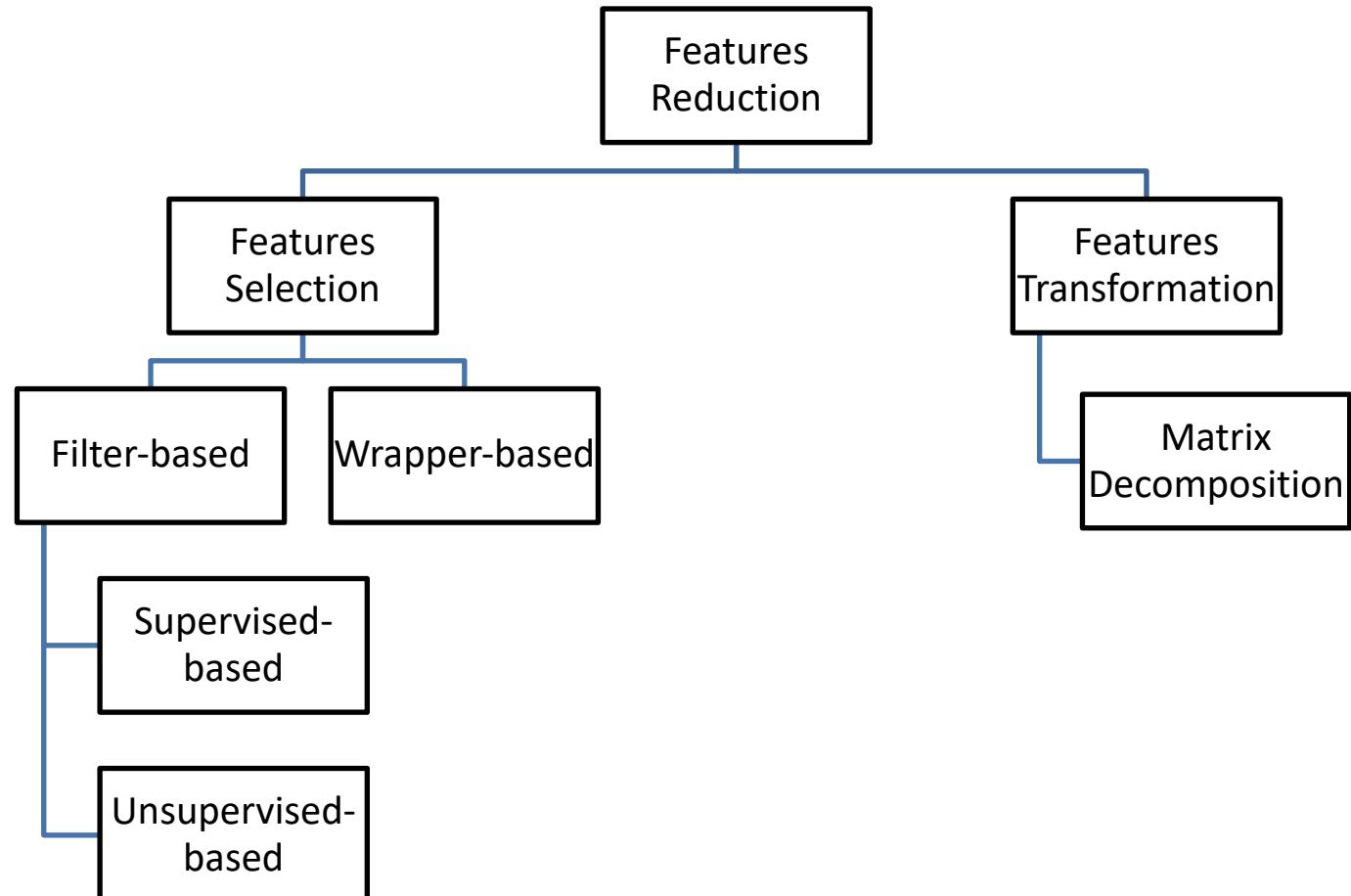


Tujuan

- ◉ Pemilihan fitur tidak hanya mengurangi dimensi ruang fitur yang tinggi, namun juga memberikan pemahaman data yang lebih baik, yang meningkatkan hasil pengelompokan.
- ◉ Penghapusan kata kunci adalah bentuk pilihan fitur.



Pengurangan Fitur





Metode Seleksi Fitur

◉ Pemilihan Fitur Berbasis Filter

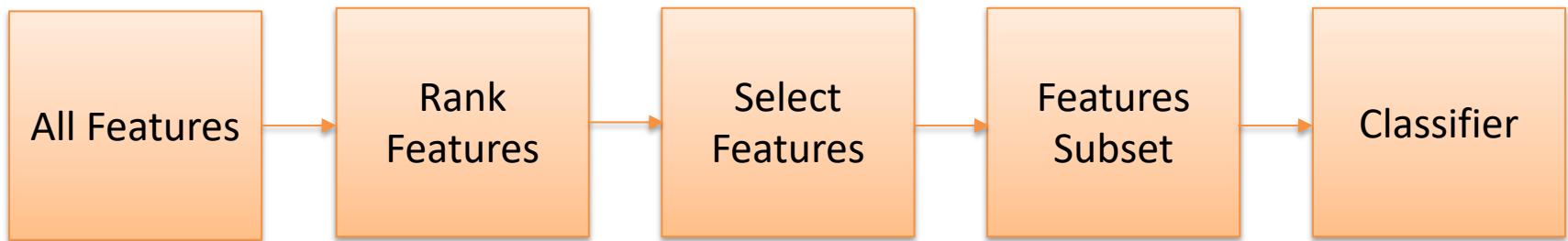
- Seleksi Fitur Supervised
- Seleksi Fitur Unsupervised

◉ Seleksi Fitur Berbasis Wrapper



Pemilihan Fitur Berbasis Filter

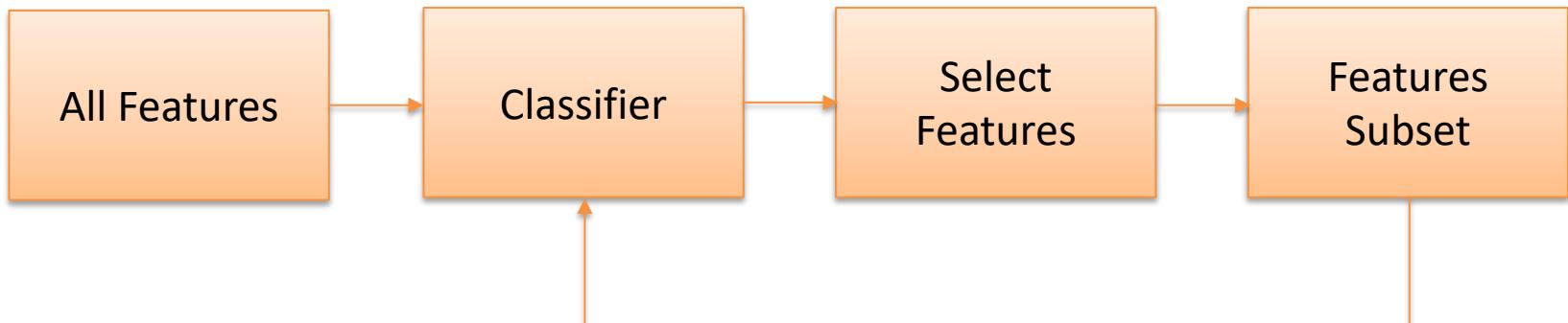
- Pilih fitur sebelum menggunakannya dalam classifier.
- Fitur akan diberi peringkat atau Bobot.





Seleksi Fitur Berbasis Wrapper

- Pilih fitur berdasarkan seberapa baik mereka bekerja dalam classifier.
- Classifier adalah bagian dari metode pemilihan fitur.
- Seringkali proses iteratif.
- Performanya tergantung pada classifier.
- Biaya komputasi tinggi.





Seleksi Fitur Supervised

- ◉ Cocok untuk klasifikasi teks.
- ◉ Membutuhkan kelas dokumen untuk menyimpan fitur yang relevan.
- ◉ Metode:
 - Chi Square
 - Information Gain

Sumber: Yang dan Pederson, "Studi Perbandingan Seleksi Fitur dalam Kategorisasi Teks"



Seleksi Fitur Supervised- Chi Square

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}.$$

$$\chi_{avg}^2(t) = \sum_{i=1}^m P_r(c_i) \chi^2(t, c_i)$$

$$\chi_{max}^2(t) = \max_{i=1}^m \{\chi^2(t, c_i)\}$$



Seleksi Fitur Supervised- Information Gain

$$\begin{aligned} G(t) = & - \sum_{i=1}^m P_r(c_i) \log P_r(c_i) \\ & + P_r(t) \sum_{i=1}^m P_r(c_i|t) \log P_r(c_i|t) \\ & + P_r(\bar{t}) \sum_{i=1}^m P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}) \end{aligned}$$



Seleksi Fitur Unsupervised

- Cocok untuk Clustering Text.
- Metode:
 - Document Frequency (DF)
 - Term Contribution (TC)
 - Term Variance Quality (TVQ)
 - Term Variance (TV)

Dari: Luying LIU, et. Al. "Sebuah Studi Banding tentang Metode Seleksi Fitur yang Tidak Disarankan untuk Klaster Teks"



Seleksi Fitur Unsupervised – Document Frequency (DF)

- Metode sederhana.
- Fitur / atribut yang ada dalam beberapa dokumen dianggap tidak penting.



Seleksi Fitur Unsupervised -TC

$$f(t_i, D_j) = TF_{ij} * \log\left(\frac{N}{DF_j}\right)$$

$$TC(t_k) = \sum_{i, j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j)$$



Seleksi Fitur Unsupervised - TVQ

$$q(t_i) = \sum_{j=1}^n f_{ij}^2 - \frac{1}{n} \left[\sum_{j=1}^n f_{ij} \right]^2$$



Seleksi Fitur Unsupervised - TV

$$v(t_i) = \sum_{j=1}^N [f_{ij} - \bar{f}_i]$$



Paper : ”A Two-stage Feature Selection Method for Text Categorization”

- Diusulkan oleh Jiana Meng dan Hongfei Lin.
- Konferensi Ketujuh Internasional tentang Fuzzy Systems and Knowledge Discovery (FSKD 2010).
- Mereka menggabungkan Feature-based method dan semantic-based method untuk mengurangi ruang vektor (term-document).



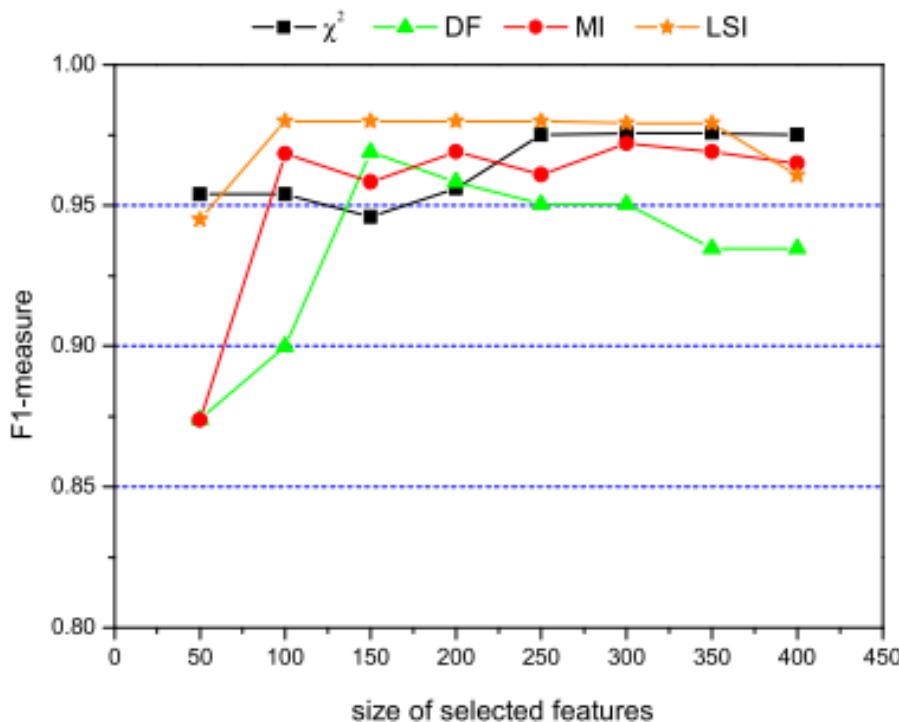
Dataset : Kumpulan Data Spam

- Dua kumpulan data tersebut mewakili pengumpulan pesan email termasuk teks header, subjek, dan body tanpa perubahan apapun, dikonversi ke format vektor fitur yang menunjukkan jumlah kejadian fitur (kata) tertentu dalam pesan.

Dataset	Ham (%)	Spam (%)	Total
LingSpam	83.4	16.6	2893
Andrew Farrugia	49.75	50.25	5243



Pemilihan fitur dan kinerja pada corpus LingSpam



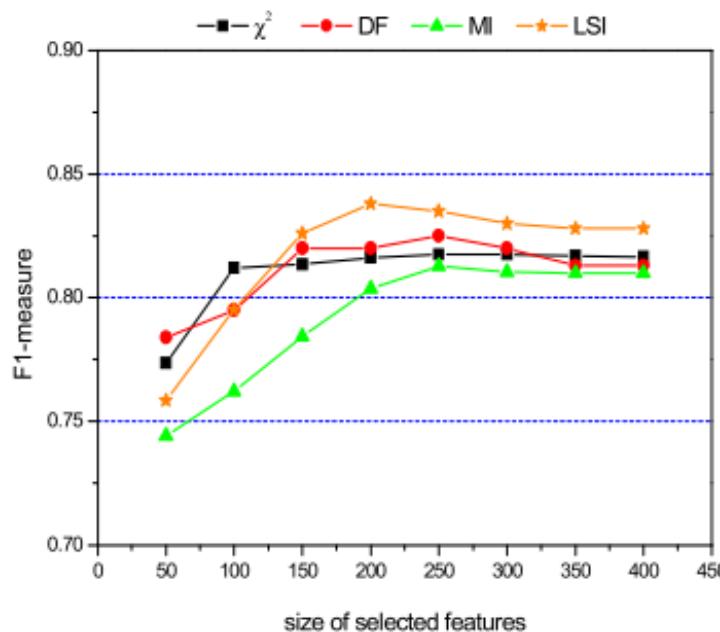
Two-stage method Performance

	k=25	k=50	k=75	k=100	k=125
F1-measure	0.9792	0.9897	0.9362	0.98	0.9691
Accuracy	0.9828	0.9897	0.9794	0.9828	0.9759
AUC	0.9987	0.9997	0.9997	0.9998	0.9991

Nilai F1-measure terbaik adalah 0,98



Pemilihan fitur dan performa pada corpus Andrew Farrugia



Pertunjukan metode dua tahap

	k=25	k=50	k=75	k=100	k=125
F1-measure	0.8519	0.8573	0.8626	0.8855	0.9691
Accuracy	0.8195	0.8219	0.8354	0.8382	0.8408
AUC	0.9513	0.9572	0.9578	0.9597	0.9578

Nilai F1-measure terbaik adalah 0.838.



- **Kesimpulan & Review**

Pemilihan fitur tidak hanya mengurangi dimensi ruang fitur yang tinggi, namun juga memberikan pemahaman data yang lebih baik, yang meningkatkan hasil pengelompokan. Penghapusan kata kunci adalah bentuk pilihan fitur.

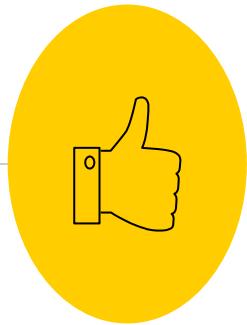
Terdapat dua metode pemilihan fitur yaitu :

1. Pemilihan Fitur Berbasis Filter.
2. Seleksi Fitur Berbasis Wrapper.



Kuis (Latihan Soal)

1. Apa itu Feature Selection untuk klasifikasi?
2. Mengapa Feature Selection itu penting?
3. Apa itu metode filter dan bagaimana pendekatan wrapper terhadap Feature Selection ?
4. Carilah minimal 2 buah (Paper/jurnal) yang membahas tentang feature selection, lakukan review dan simpulkan bagaimana peran feature selection pada penelitian yang dilakukan.



Thanks!

Any *questions* ?