MODUL DATA MINING

TEXT MINING
PERTEMUAN 12 (ONLINE)

Disusun Oleh
Syefira Salsabila

Dalam ilmu Data Mining di kenal dengan istilah Text Mining. Text mining atau text analytics adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan data mining dimana data mining mengolah data yang sifatnya terstruktur. Text mining adalah sebuah penelitian baru yang menarik yang mencoba memecahkan masalah informasi yang overload, dengan menggunakan teknik dari data mining, machine learning, natural language processing (NLP), information retrieval (IR), dan knowledge management. Text mining melibatkan preprocessing koleksi dokumen (kategorisasi teks, ekstraksi informasi, ekstraksi istilah), penyimpanan representasi menengah, teknik untuk menganalisis representasi menengah ini (seperti analisis distribusi, pengelompokan, analisis tren, dan peraturan asosiasi), dan visualisasi hasilnya.

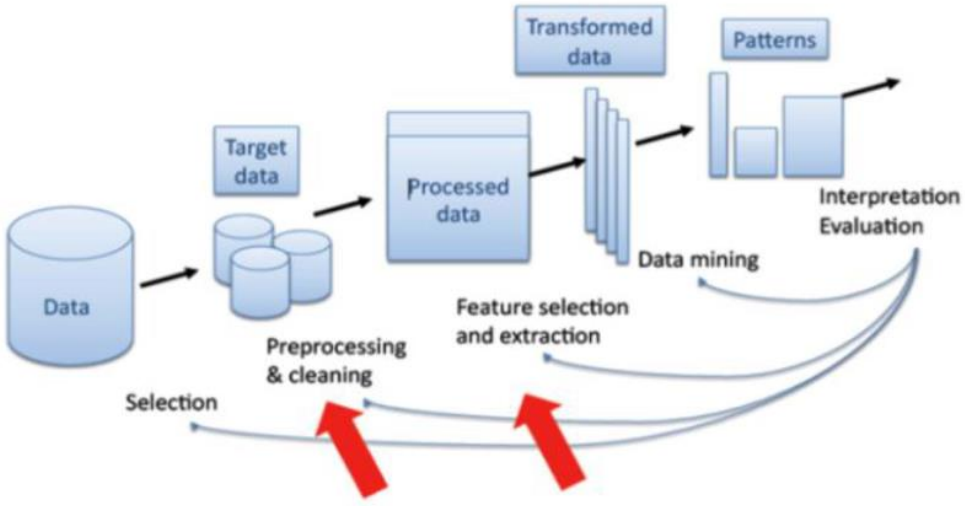
Dalam melakukan text mining, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses *text preprocessing*. *Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur.

Teks yang dilakukan proses text mining pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat noise pada data, dan terdapat struktur teks yang tidak baik. Cara yang digunakan dalam mempelajari struktur data teks adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen, sebelum menentukan fitur-fitur yang mewakili, diperlukan tahap pre-processing. Tujuan utama text preprocessing adalah untuk mendapatkan bentuk data siap olah untuk diproses oleh data mining dari data awal yang berupa data tekstual. Adapun tahap-tahap text preprocessing yang dilakukan adalah sebagai berikut:

1. Case folding, merupakan proses pengubahan huruf dalam dokumen menjadi satu bentuk, misalnya huruf kapital menjadi huruf kecil dan sebaliknya.
2. Tokenizing, merupakan proses pemisahan teks menjadi potongan kalimat dan kata yang disebut token.
3. Filtering, merupakan proses membuang kata-kata serta tanda-tanda yang tidak bermakna secara signifikan, seperti hashtag (#), url, tanda baca tertentu (emoticon), dan lainnya.
4. Stemming, merupakan proses pengubahan kata ke dalam bentuk kata dasar, sehingga berfungsi mengurangi jumlah indeks yang berbeda dari suatu dokumen.

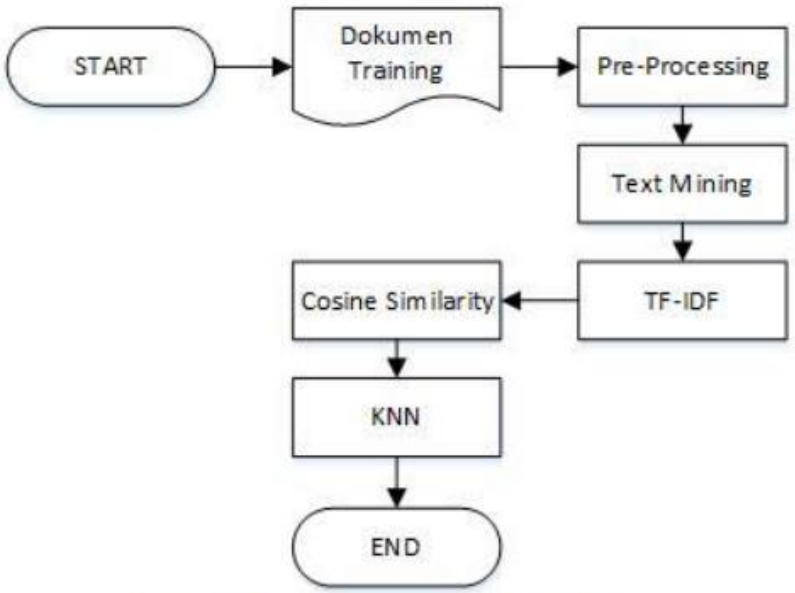
Dalam bidang komputerisasi yang termasuk kedalam *machine learning*, *Naïve Bayes* dan *Support Vector Machine* (SVM) merupakan metode yang digunakan untuk klasifikasi teks dalam *text mining*. Sebagai salah satu metode komputasi yang efisien dan mempunyai *performance preditictive* yang baik, *naïve bayes* merupakan salah satu metode klasifikasi teks yang popular. *Naïve Bayes* merupakan algoritme yang sering

digunakan dalam pengkategorian teks, dimana konsep dasarnya adalah menggabungkan probabilitas kata-kata dan kategori sebuah dokumen.



Gambar 7: Proses *Data Mining*

Pada **Tahap Pertama**: Melakukan persiapan data dalam dari log sistem yang ada dalam aplikasi server pemberitaan. Data dipilih dari sekian banyak data log khusus bagian log transaksi sistem sms.
Tahap Kedua: Melakukan pre-prosesing dengan text mining yang meliputi *tokenizing, filtering, steaming* dan *stopward* sehingga data siap diolah ke proses berikutnya.
Tahap Ketiga: Melakukan pengolahan data dengan menggunakan algoritma data mining yang terdiri dari, estimasi, prediksi, klasifikasi, cluster dan asosiasi.
Tahap Keempat menentukan dictionary manual terkait dengan pemisahan content isi SMS terdiri dari Label Kerja dan tidak Kerja.
Tahap Kelima: Melakukan evaluasi komparasi dari hasil untuk mengukur tingkat akurasi kerja proses text mining dengan KNN terhadap proses manualnya.



Gambar 1. Proses Text *Mining*

Pada tahap pengolahan text *mining* dilakukan tahap *preprocessing* yang dijelaskan pada Gambar 2.

Gambar 2. *Preprocessing* Text *Mining*

Text mining adalah penggalian informasi dari teks oleh user menggunakan *tools* analisis. Secara umum *text mining* mengadopsi proses-proses didalam data *mining* dan didalam *text mining* juga menggunakan teknik data *mining*.

Text preprocessing menjadi tahap awal dalam *text mining*. *Preprocessing* dilakukan untuk menghilangkan bagian atau teks yang tidak diperlukan sehingga mendapatkan data yang berkualitas untuk dieksekusi. Pertama dalam tahap *preprocessing* yaitu *tokenizing* yang bertujuan untuk memecah kalimat menjadi perkata yang terpisah dikenal dengan nama *term* atau token. Selanjutnya *filtering* dengan melakukan penghapusan tanda baca, merubah huruf capital menjadi huruf kecil dan penghapusan *stopword* yang bertujuan untuk menghapus kata-kata yang tidak bermanfaat atau tidak memiliki pengaruh dalam proses. Terakhir yaitu *stemming* untuk mendapatkan kata dasar dari kata yang telah mendapatkan imbuhan atau keterangan lainnya. *Stemming* yang digunakan yaitu *stemming* Nazief dan Adriani karena Algoritma ini memiliki akurasi lebih besar dibandingkan dengan algoritma porter.

Dengan perkembangan teknologi yang semakin besar maka kebutuhan akan penyajian informasi yang cepat dan akurat menjadi salah satu focus utama dalam

penelitaian dan pengembangan guna memenuhi kebutuhan informasi yang semakin cepat dan akurat. Data Mining merupakan kompleks teknologi yang berakar pada berbagai disiplin ilmu: matematika, statistik, ilmu komputer, fisika, teknik, biologi, dll, dan dengan beragam aplikasi dalam berbagai macam domain yang berbeda: bisnis, kesehatan, sains dan teknik , dll Pada dasarnya, data mining dapat dilihat sebagai ilmu menjelajahi dataset besar untuk mengekstraksi informasi tersirat, yang sebelumnya tidak diketahui dan berpotensi berguna.

Sedangkan *Text mining* adalah salah satu penambangan informasi yang berguna dari data – data yang berupa tulisan, dokumen atau text dalam bentuk klasifikasi maupun clustering. Text mining masih merupakan bagian dari data mining dimana akan memproses data – data atau text – text serta dokumen – dokumen yang bisa jadi dalam jumlah sangat besar. Untuk memproses data yang sangat besar tentulah akan memakan sumber daya yang tidak sedikit kaitanya dengan pengolahan data tersebut. Disinilah diperukanya sebuah pemrosesan awal atau preprocessing data text tersebut sebelum data tersebut di lakukan proses text mining sesuai algoritma yang akan diterapkan.

Dengan *text mining* maka kita akan melakukan proses mencari atau penggalian informasi yang berguna dari data tekstual. Ini juga merupakan salah satu kajian penelitian yang sangat menarik dan juga sangat berguna di kemudian hari dimana seperti mencoba untuk menemukan pengetahuan dari dokumen–dokumen atau teks - teks yang tidak terstruktur. *Text mining* sekarang juga memiliki peran yang semakin penting dalam negara berkembangaplikasi, seperti mengetahui isi dari teks secara langsung dari proses *text mining* tanpa perlu membaca satu persatu teks atau tulisan yang ada. Proses Text mining adalah sama dengan data mining, kecuali, beberapa metode dan data yang di kelola nya seperti data teks yang tidak terstruktur, terstruktur sebagian maupun terstruktur seperti teks email, teks HTML, maupun teks komentar serta dari berbagai sumber.

Untuk dapat melakukan penambangan informasi atau text mining maka perlu dilakukan beberapa tahapan yang harus dilakukan untuk mengolah sumber data baik yang terstruktur, terstuktur sebagian dan yang tidak terstruktur dari beberapa sumber maka data-data tersebut perlu dilakukan proses awal atau di sebut sebagai preprocessing text yang bermaksud mengolah data awal yang masih bermacam – macam untuk dijadikan sebuah data teratur yang dapat dikenai atau diterapkan beberapa metode text mining yang ada.

Apa sih arti text mining yang sebenarnya? Definisi akan text mining sudah sering di berikan oleh banyak ahli riset dan praktisi. Seperti hal nya data mining, text mining adalah proses penemuan akan informasi atau trend baru yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan unstructured text, text mining mencoba untuk mengasosiasikan satu bagian text dengan yang lainnya berdasarkan aturanaturan tertentu. Hasil yang di harapkan adalah informasi baru atau “insight” yang tidak terungkap jelas sebelumnya. Seperti halnya data mining, text mining juga menghadapi

masalah yang sama, termasuk jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data “noise.” Berbeda dengan data mining yang utamanya memproses structured data, data yang digunakan text mining pada umumnya dalam bentuk unstructured, atau minimal semistructured, text. Akibatnya, text mining mempunyai tantangan tambahan yang tidak di temui di data mining, seperti struktur text yang complex dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah translasi yang tidak akurat.

Dikarenakan structured data ditujukan agar mudah di proses komputer secara automatic, pre-process data di data mining jauh lebih mudah dilakukan dari pada pada unstructured text. Text di ciptakan bukan untuk di gunakan oleh mesin, tapi untuk dikonsumsi manusia langsung. Karena itu, pada umumnya “Natural Language Processor” digunakan untuk memproses unstructured text. Hearst mempertanyakan pengunaan kata ‘mining’ di data mining dan text mining. Kata ‘mining’ memberikan arti dimana fakta-fakta atau relasirelasi baru dihasilkan dari proses me-‘mining’ data. Dia mengklaim bahwa aktivitas data mining lebih memfocuskan pada penemuan trend dan pattern yang sebenarnya sudah ada. Sedangkan ahli text mining yang lain beranggapan bahwa text mining adalah proses penemuan kembali relasi dan fakta yang terkubur didalam text, dan tidak harus baru.

Ulasan di berikutnya sedikit mengikuti definisi text mining oleh Hearst. Seperti di sebutkan sebelumnya, Text mining telah mengadopsi teknik yang di gunakan di bidang natural language processing dan computational linguistics. Walaupun teknik di computational linguistics bisa dibilang maju dan cukup akurat untuk mengekstrak informasi, tujuan text mining bukan hanya mengekstrak informasi. Melainkan untuk menemukan pattern dan informasi baru yang belum terungkapkan, yang sulit ditemukan tanpa analisa yang dalam. Walau kemampuan komputer untuk mencapai kemampuan untuk memproses text seperti manusia sangat sulit, bila tidak mustahil, telah banyak teknik-teknik baru di computational linguistics yang bisa membantu text mining untuk mencerna text lebih jauh lagi.

Sering kali penguna search engine di Internet menganggap search engine sebagai salah satu implementasi text mining. Andil utama search engine hanyalah menyingkirkan text yang tidak memiliki kata-kunci yang di cari pengunna. Dan lagi penguna search engine mengetahui sebelumnya text seperti apa yang hendak dia cari. Bisa dibilang kalau pencarian seperti ini termasuk dalam “Information Retrieval.” Focus information retreieval adalah menemukan dokumen atau text yang memenuhi kriteria pencari. Text mining lebih memfokuskan pada relasi dan co-existence dari satu dokumen dengan yang lainnya. Walaupun text mining lebih dari information retrieval, text mining telah mengadopsi information retrieval untuk menyaring dan mengurangi jumlah informasi untuk diproses selanjutnya. Methode statistik juga sudah mulai sering di gunakan dan di adopsi di computational linguistics dan information retrieval yang nanti nya bisa memberikan tool yang lebih baik dan akurat untuk text mining.

Banyak juga ahli riset yang menkategorikan document categorization sebagai text mining. Walau kategorisasi dokumen dapat memberikan label dan kesimpulan yang

akurat pada dokumen-dokumen tertentu, ini tidak menghasilkan fakta-fakta atau relasi yang baru. Tetapi bilamana label-label atau kesimpulankesimpulan yang di hasilkan di analisa dan di korelasikan lebih lanjut, ini bisa menghasilkan fakta dan relasi baru antara group-group dokumen yang berbeda. Kegiatan seperti ini bisa di masukan dalam text mining.

Contoh dalam mengolah data teks ulasan sebuah hotel dalam bahasa Indonesia yang masih memiliki bentuk tidak terstruktur ke dalam bentuk yang lebih terstruktur dan melakukan ekstraksi informasi dari sejumlah ulasan dengan membuat sebuah *wordcloud. Wordcloud* merupakan kumpulan kata-kata yang paling sering dibicarakan dalam ulasan.

Menggunakan software R Programming untuk melakukan analisis teks, sebelum masuk pada tahap pembuatan *wordcloud,* terlebih dahulu saya akan melakukan *text preprocessing* untuk cleaning data agar data siap di olah. Untuk melakukan analisis teks, dalam R digunakan beberapa *packages* diantaranya *packages "tm", "RColorBrewer", "wordcloud"* dan *"stringr" .* Untuk menginstall packages tersebut ke dalam program R, dapat dilakukan dengan cara menjalankan script berikut:

```
install.packages("tm")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("stringr")
```

Aktifkan packages dengan perintah "library"

```
library("tm")
library("RColorBrewer")
library("wordcloud")

library("stringr")
```

Kemudian, aktifkan folder kerja yang merupakan tempat penyimpanan file postingan (dalam bentuk .csv) dengan perintah "setwd" dan baca file ke dalam R menggunakan perintah "readLines" seperti berikut:

```
setwd("E://KULIAH")
docs<-readLines("dataulasan.csv")
docs
```

Sehingga akan tampil isi ulasan seperti berikut:

```
> docs
 [1] "\"Kamar hotel cukup luas, bersih, dan nyaman.. memiliki view dan teras mengh$
 [2] "kesini sekitar tahun 2013 hotel cukup bagus kamar luas dan nyamansarapan len$
 [3] "\"Hotelnya cukup tenang, nyaman untuk beristirahat. Pilihan menu sarapannya $
 [4] "\"Menyenangkan, lokasi yang strategis, dekat mall, kuliner, kolam renang yan$
 [5] "terakhir memasuki royal ambarukmo adalah saat masih bernama ambarukmo.. tahu$
 [6] "\"Business and Leisure combines together.Staff hotel nya ramah-ramah dan tan$
 [7] "\"Sarapan pagi lengkap and enak, kolam renang and gymnya bersih dan ok, suas$
 [8] "\"Saya puas dengan segala fasilitas yang ada disini, sangat memuaskan. Pool $
 [9] "ketika anda di yogya dan bingung mau nginep dimana...Hotel ini menyediakan a$
[10] "Reception kurang profesional tidak mengerti membaca Remark apa yang di jual $
[11] "\"Hotelnya bagus, bersih, sangat nyaman buat keluarga menginap. Staffnya jug$
[12] "\"Satu kata untuk hotel ini yaitu terbaik. Hotel yang asri, banyak tanaman a$
[13] "\"Royal Ambarukmo mempunyai sejarah yang menarik,bahkan sebagian dari bangun$
[14] "Hotel yang memberikan pelayanan breakfast yang enak dengan variasi yang bera$
```

Data ulasan di atas masih berbentuk tidak terstruktur, dan masih banyak noise sehingga perlu dilakukan cleaning data. Untuk melakukan cleaning data terlebih dahulu data teks harus di ubah ke dalam bentuk Corpus dengan menjalankan script berikut:

```
docs <- Corpus(VectorSource(docs))
```

Kemudian, akan dilakukan pembersihan data, dengan mengganti tanda “/”, “@” and “|” dengan sebuah spasi menggunakan perintah:

```
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

Kemudian dilakukan proses *case folding,* yakni menyeragamkan huruf ke dalam bentuk huruf kecil menggunakan perintah

```
docs <- tm_map(docs, content_transformer(tolower))
```

Kemudian menghapus tanda baca (punctuation) dengan menggunakan perintah:

```
docs <- tm_map(docs, toSpace, "[[:punct:]]")
```

Menghapus angka dengan menggunakan perintah:

```
docs <- tm_map(docs, toSpace, "[[:digit:]]")
```

Kemudian dilakukan proses filtering yakni membuang daftar kata-kata yang kurang penting untuk di analisis menggunakan stopwords. Stopword / stoplist adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Untuk

menjalankan stopwords dapat dilakukan dengan menjalankan perintah berikut:

```
myStopwords = readLines("stopword_id.csv")
docs <- tm_map(docs, removeWords, myStopwords)
```

Untuk menghapus daftar kata secara manual juga dapat dilakukan dengan cara berikut:

```
docs <- tm_map(docs, removeWords, c("you","also","hotel","ambarrukmo","royal"))
```

Menghapus spasi yang tidak berguna, yakni terdapat spasi yang berlebih pada selah antara dua kata, untuk menghapus spasi berlebih tersebut dapat digunakan perintah berikut:

```
docs <- tm_map(docs, stripWhitespace)
```

Menghapus URL web dengan menjalankan perintah:

```
removeURL <- function(x) gsub("http[[:alnum:]]*", " ", x)
docs <- tm_map(docs, removeURL)
```

Untuk memperbaiki kata-kata yang salah (spelling normalization), dapat dilakukan dengan cara manual menggunakan perintah:

```
docs <- tm_map(docs, gsub, pattern="Howver", replacement="However")
docs <- tm_map(docs, gsub, pattern="goood", replacement="good")
```

Setelah melalui tahap cleaning data, kemudian merubah data ke dalam bentuk Term Document Matrix, dan mengubah ke dalam bentuk data frame sehingga dapat dihitung frekuensi setiap kata. Adapun perintah yang digunakan adalah sebagai berikut:

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

Sehingga akan tampil jumlah frekuensi kata seperti berikut:

```
> dtm <- TermDocumentMatrix(docs)
> m <- as.matrix(dtm)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
> head(d, 10)
        word freq
kolam   kolam   12
kamar   kamar    8
luas     luas    8
renang renang    8
nyaman nyaman    7
mall     mall    6
enak     enak    6
bagus   bagus    5
ramah   ramah    5
and       and    5
```

Setelah di peroleh frekuensi setiap kata, kemudian kata-kata tersebut dapat di ubah ke dalam bentuk wordcloud menggunakan perintah berikut:

```
dtm <- TermDocumentMatrix(docs)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=50, random.order=FALSE, rot.per=0.35,

    colors=brewer.pal(8, "Dark2"))
```

Sehingga akan muncul tampilan wordcloud seperti gambar berikut:

Berdasarkan gambar wordcloud diatas, semakin besar kata pada tampilan wordcloud maka menunjukkan semakin besar pula frekuensi kata tersebut dalam dokumen ulasan.

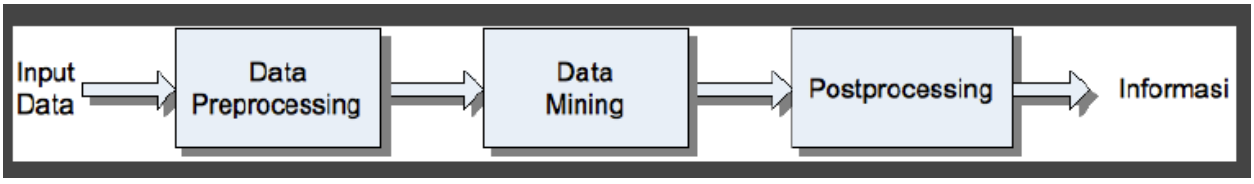
Aplikasi text mining

Aplikasi text mining bisa di bagi berdasarkan tipe unstructured text yang di proses. Untuk ustructured text dalam bentuk emails, intant messages, dan blogs, pada umumnya pengguna ingin mencari atau “mine” informasi mengenai orang (seperti email pengirim, alamat, nama lengkap, dll), perusahaan (seperti nama lengkap dan lokasi), organisasi, dan kejadian-kejadian (seperti penemuan baru, pengumuman penting, dll). Untuk berita dari berbagai sumber, text mining bisa di gunakan untuk membandingkan berita yang sama atau berbeda yang berasal dari sumber yang berbeda, mungkin dengan bahasa yang berbeda. Lebih jauh lagi adalah analisa dan organisasi isi berita berdasarkan waktu publikasi (atau “temporal analysis”). Text mining juga bisa membantu untuk proses “deduplication” di sini. Untuk buku-buku dan artikel-artikel science, text mining di butuhkan untuk mendeteksi trend di bidang riset tertentu. Salah satu cara yang bisa di lakukan adalah dengan memonitor jumlah publikasi untuk bidang riset tertentu untuk jangka waktu tertentu. Hasil-hasil untuk bidang riset yang berbeda bisa di bandingkan dan di analisa guna memberikan hasil trend yang berarti. Untuk technical working paper, documentasi, dan software specifikasi dokumen, text mining bisa di gunakan untuk mengekstrak software requirement dari spesifikasi dokumen secara otomatis atau mendeteksi ke kurangan antara source code dan documentasinya secara otomatis. For web pages, text mining bisa di gunakan untuk menganalisa website perusahaan, structur websitenya, perbandingan website content yang satu dengan site yang lain. Masih banyak lagi aplikasi text mining yang di butuhkan.

Proses Text Mining

Proses text mining mencakup beberapa sub-task, seperti information retrieval, categorization, POS tagging, Clustering, dan lainnya, yang bisa di katergorikan kedalam framework “Knowledge Discovery in Databases” (KDD), yang tidak lain adalah proses mengidentifikasikan pattern di dalam data yang benar, unik, berguna, dan dimengerti. KDD proses interaktif, bisa berulang, dan terdiri dari step Selection, Preprocessing, Transformation, Data Mining, dan Interpretation/Evaluation. Dalam sesi ini, proses dan kegiatan text mining yang beragam akan saya coba assosiasikan dengan KDD step dan ulas secara singkat.

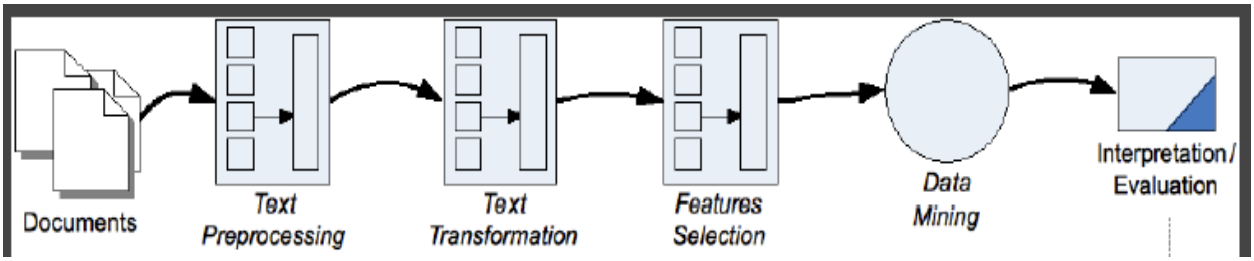
Data mining adalah suatu proses yang secara otomatis mencari atau **menemukan informasi** yang bermanfaat dan suatu kumpulan data yang besar. Data Mining lebih dekat pada bidang **pencarian pengetahuan** dalam basis data (knowledge discovery in database / KDD), yang merupakan proses **konversi** dari data mentah menjadi informasi yang bermanfaat.



Data mining dibagi dalam dua kelompok jenis tugas analisis data:

a. Predictive task : bertugas untuk **memprediksi nilai** sebuah atribut tertentu (target) didasarkan pada nilai atribut lain (*explanatory)*
b. Descriptive task : bertugas **mendapatkan pola** analisis asosiasi (*association analysis), pengelompokan (clustering),* penyimpangan (*anomaly detection) yang meringkas* hubungan-hubungan dalam data

Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks, yaitu proses penganalisisan teks guna menyarikan informasi yang bermanfaat untuk tujuan tertentu. Berdasarkan ketidakteraturan struktur data teks, maka proses text mining memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur.



Tahapan Text Mining

Text mining merupakan penerapan konsep dan teknik data mining untuk **mencari pola dalam teks**. Teks Mining : Proses penganalisisan teks guna **menyarikan informasi** yang bermanfaat untuk tujuan tertentu.

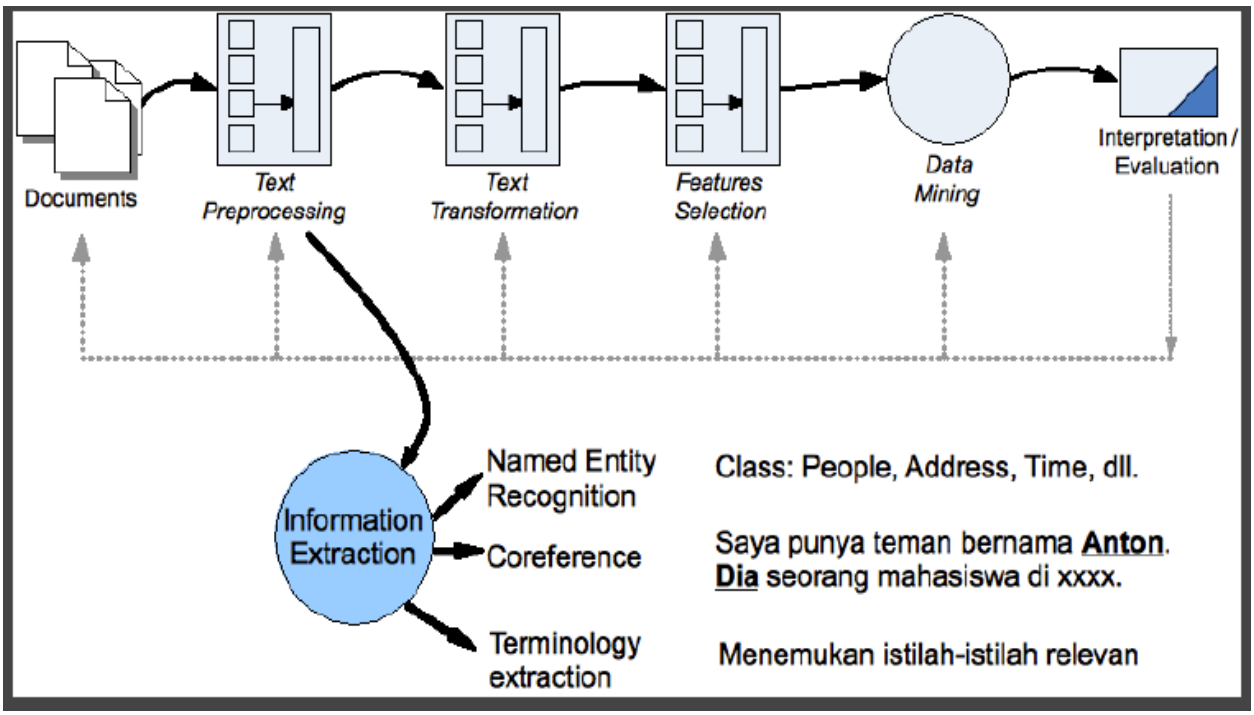
Perbedaan mendasar dengan Data Mining pada umumnya, **Text Mining** mengolah data **teks yang tidak terstruktur**, maka proses text mining memerlukan beberapa tahap awal ***(preprocessing)*** yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur.

Perbedaan	Data Mining	Text Mining
Data Object	Numerical & categorical data	Textual data
Data structure	Structured	Unstructured & semi-structured
Data representation	Straightforward	Complex
Space dimension	< tens of thousands	> tens of thousands
Methods	Data analysis, machine learning, Neural Network, etc.	Data mining, information Retrieval, NLP, etc.
Maturity	Broad implementation since 1994	Broad implementation starting 2000

Masalah umum yang ditangani:

a. Pengorganisasian dan Clustering Dokumen
   Clustering adalah pengorganisasian kumpulan pola ke dalam cluster (kelompok-kelompok) berdasar atas kesamaannya. Pola-pola dalam suatu cluster akan memiliki kesamaan ciri/sifat daripada pola-pola dalam cluster yang lainnya. Clustering bermanfaat untuk melakukan analisis pola-pola yang ada, mengelompokkan, dan membuat keputusan. Metodologi clustering lebih cocok digunakan untuk eksplorasi hubungan antar data untuk membuat suatu penilaian terhadap strukturnya.

b. Klasifikasi Dokumen
   Klasifikasi adalah mengelompokkan dokumen berdasarkan data training yang sudah dilabeli. Perbedaanya dengan clustering adalah pada klasifikasi, kelas/kategorinya sudah ditentukan di awal, sedangkan pada clustering tidak.



c. Information Extraction
   Information Extraction bermanfaat untuk menggali struktur informasi dari sekumpulan dokumen. Dalam menerapkan IE, perlu sekali dilakukan pembatasan domain problem. IE sangat memerlukan NLP untuk mengetahui gramatikal dari setiap kalimat yang ada. Sebagai contoh:

a. “Indonesia dan Singapore menandatangani MoU kerjasama dalam bidang informasi dan komunikasi.”
b. KerjaSama(Indonesia, Singapore, TIK)

Dengan IE, kita dapat menemukan:

a. concepts (CLASS)
b. concept inheritance (SUBCLASS-OF)
c. concept instantiation (INSTANCE-OF)
d. properties/relations (RELATION)
e. domain and range restrictions (DOMAIN/RANGE)
f. equivalence

Web Mining

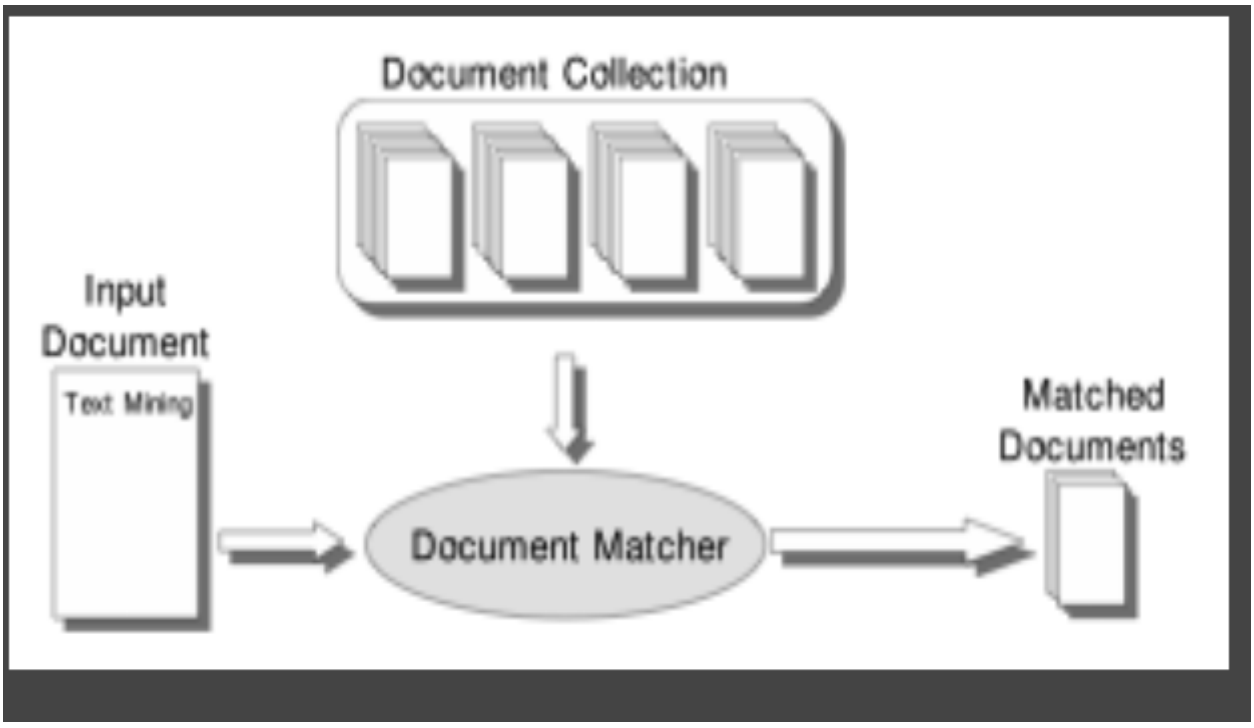
Jumlah data/informasi di web sangat besar dan terus bertambah.

a. tipe data beragam
b. informasi pada web sangat beragam.
c. informasi-informasi di web saling terhubung.
d. informasi di web sangat "kotor".
e. web juga merupakan service.
f. web dinamis
g. web merupakan sarana komunitas sosial virtual.

Web Mining bertujuan untuk menemukan informasi atau pengetahuan dari Web hyperlink structure, contoh :menemukan halaman web terpenting; menemukan komunitas pemakai yang berbagi ketertarikan topik yang sama.

Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah melakukan pengolahan untuk **memahami Bahasa alami** yang diucapkan manusia Bahasa alami adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Bahasa yang diterima oleh komputer butuh untuk diproses dan dipahami terlebih dahulu supaya maksud dari user bisa dipahami dengan baik oleh komputer.



Information Retrieval (IR)

Konsep dasar dari IR adalah pengukuran kesamaan sebuah perbandingan antara dua dokumen, mengukur sebearapa mirip keduanya. Setiap input query yang diberikan, dapat dianggap sebagai sebuah dokumen yang akan dicocokan dengan dokumendokumen lain. Pengukuran kemiripan serupa dengan metode klasifikasi yang disebut metode nearest-neighbour.

Perbedaan mendasar antara Text Mining dan IR :

a. Text Mining : Discovery of novel information
   Extracting Ore from otherwise worthless rock : menemukan informasi yang relevan dan bermanfaat dari sekumpulan data besar yang kelihatanya tidak berguna.
b. IR : Retrieval of Non-novel Information
   Finding needles in a needle-stack : mencari informasi yang relevan di antara informasi-informasi lain yang berguna namun tidak relevan

Search Engine merupakan aplikasi nyata dari **Information Retrieval** pada bidang web.

DAFTAR PUSTAKA

Hamzah, A. (2012). Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis. In *Prosiding Seminar Nasional*.

Sanjaya dan Absar. *Pengelompokkan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K-Nearest Neighbour*. Jurnal CoreIT. 2015; Vol. 1, No. 2, Desember.