BRIN
BADAN RISET
DAN INOVASI NASIONAL

Modul

PENGUMPULAN DAN PERSIAPAN DATA

Pelatihan Jabatan Fungsional Analis Data Ilmiah

Fungsi Layanan Pengembangan Kompetensi Kedinasan

Direktorat Pengembangan Kompetensi
Deputi Sumber Daya Manusia
2022

Penanggung Jawab:

1. Edy Giri Racman Putra, Ph.D.
2. Nining Setyowati Dwi Andayani, S.E., M.M.
3. Raden Arthur Ario Lelono, Ph.D.
4. Alpha Fadila Juliana Rahman, S.Pd., M.Pd.

Tim Penyusun Modul:

1. Zaenal Akbar, M.Kom., Ph.D.
2. Hendro Subagyo, M.Eng.
3. Dr. Hanif Fakhrurroja, S.SI., M.T.
4. Vera Purba Wisesa, S.Pd., M.Pd.

Diterbitkan oleh:

Direktorat Pengembangan Kompetensi - BRIN

Gedung B.J. Habibie, Jalan M.H. Thamrin Nomor 8

Jakarta Pusat 10340

Diterbitkan pertama kali tahun 2022

KATA PENGANTAR

Pengembangan kompetensi Aparatur Sipil Negara (ASN) khususnya Pegawai Negeri Sipil (PNS) dalam mengembangkan karier jabatan fungsionalnya menjadi suatu tuntutan sehingga mampu menjalankan tugas dan fungsinya dengan sebaik – baiknya seingga mampu memberikan kontribusi yang nyata terhadap bangsa dan Negara Kesatuan Republik Indonesia (NKRI).

Badan Riset dan Inovasi Nasional (BRIN) sebagai Instansi pembina 11 (sebelas) jabatan fungsional yang meliputi peneliti, perekayasa, pengembangan teknologi nuklir, analis perkebunrayaan, analis pemanfaatan iptek, analis data ilmiah, kurator koleksi hayati, penata penerbitan ilmiah, teknisi perkebunrayaan, teknisi penelitian dan perekayasaan, dan pranata nuklir. BRIN melalui kedeputian Sumber Daya Manusia Ilmu Pengetauan dan Teknologi (SDMI) bertanggung jawab dalam penyelenggaraan pengembangan kompetensi 11 (sebelas) jabatan fungsional tersebut.

Kedeputian SDMI – BRIN melalui Direktorat Pengembangan Kompetensi sebagai penyelenggara pengembangan komptensi jabatan fungsional bertanggung jawab dalam menyiapkan kebutuhan tersebut baik berupa pengelolaan pembelajaran, fasilitator, modul, bahan ajar dan sebagainya, yang merujuk kepada regulasi peraturan BRIN nomor 28 tahun 2022 tentang pedoman pelatihan pembentukan jabatan fungsional peneliti, peraturan BRIN nomor 29 tahun 2022 tentang pedoman pelatihan jabatan fungsional kurator koleksi hayati, peraturan BRIN nomor 30 tahun 2022 tentang pedoman pelatihan jabatan fungsional analis pemanfaatan iptek, peraturan BRIN nomor 31 tahun 2022 tentang pedoman pelatihan jabatan fungsionalanalis data ilmiah, peraturan BRIN nomor 32 tahun 2022 tentang pedoman pelatihan jabatan fungsional analis perkebunrayaan, peraturan BRIN nomor 33 tahun 2022 tentang pedoman pelatihan jabatan fungsional teknisi perkebunrayaan, dan peraturan BRIN nomor 34 tahun 2022 tentang pedoman pelatihan jabatan fungsional penata penerbitan ilmiah.

Kami megucapkan syukur ke hadirat Tuhan Yang Maha Esa, atas berkat rahmat-Nya, modul pelatihan jabatan fungsional analis data ilmiah yang berjudul

“Pengumpulan dan Persiapan Data” dapat diselesaikan tepat waktu. Modul ini digunakan dalam pelatihan jabatan fungsional yang dibina oleh BRIN yang diselenggarakan oleh Kedeputian SDMI - BRIN melalui Direktorat Pengembangan Kompetensi. Kami berharap modul ini dapat memberikan manfaat dan kontribusi dalam meningkatkan dan mengembangkan kompetensi jabatan fungsional yang dibina BRIN.

Jakarta, Desember 2022

Plt. Deputi Sumber Daya Manusia
Ilmu Pengetauan dan Teknologi
Badan Riset dan Inovasi Nasional

(*Tanda tangan*)

Edy Giri Rachman Putra, Ph.D.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

PENDAHULUAN

A. Deskripsi Singkat

Mata Pelatihan ini menjelaskan tentang mekanisme pengumpulan data yang berasal dari beberapa sumber dan mempersiapkan data tersebut untuk diolah lebih lanjut.

B. Alokasi Waktu

Pelatihan Jabatan Fungsional Analis Data Ilmiah dapat diselenggarakan dengan tiga metode. Setiap metode memiliki alokasi waktu yang berbeda. Berikut adalah alokasi waktu pembelajaran untuk mata pelatihan Pengumpulan dan Persiapan Data.

Metode	*Synchronous*	*Asynchronous*	Total
Klasikal	12 JP	-	12 JP
Bauran	8 + 3 JP	3 JP	15 JP
Jarak Jauh	12 JP	3 JP	15 JP

C. Tujuan Pembelajaran

1. Hasil Belajar

Peserta mampu mengetahui dan melakukan berbagai metode pengumpulan dan persiapan data dengan benar.

2. Indikator Hasil Belajar

Setelah selesai pembelajaran diharapkan peserta mampu:

a. Menjelaskan akusisi berbagai jenis data dari berbagai sumber dengan benar
b. Menjelaskan pentingnya label data untuk pengolahan data secara otomatis dengan benar
c. Melakukan pra-pemrosesan data seperti standarisasi, normalisasi, smoothing, dan transformasi untuk pengolahan data dengan benar
d. Mengidentifikasi fitur data untuk pengolahan data dengan benar

D. Materi Pokok

Mata pelatihan ini terdiri dari4 Materi Pokok, yaitu:

1. Akusisi data
    a. Definisi dan ruang lingkup
    b. Analisis kebutuhan/ potensi data
    c. Analisis Sumber data
    d. Pengumpulan data
2. Pelabelan Data
    a. Definisi dan ruang lingkup
    b. Metode pelabelan data
3. Prapemrosesan Data
    a. Pengenalan prapemrosesan data
    b. Pentingnya kualitas Data
    c. Validitas Data
    d. Penanganan Data yang Hilang (*missing Value*)
    e. Standarisasi/Normalisasi Data
4. Data Outlier/Deteksi Outlier Perekayasaan Fitur
    a. Definisi dan Ruang lingkup
    b. Memahami Fitur
    c. Pengembangan Fitur: Data Kategorikal
    d. Pengembangan Fitur: Data Tekstual
    e. Seleksi Fitur

MATERI POKOK 1:

AKUISISI DATA

Indikator Hasil Belajar:

Peserta mampu **menjelaskan akuisisi berbagai jenis data dari berbagai sumber dengan benar**, meliputi:

1. Menjelaskan definisi dan ruang lingkup akuisisi berbagai jenis data dari berbagai sumber dengan benar
2. Menjelaskan Analisis kebutuhan/ potensi data akuisisi berbagai jenis data dari berbagai sumber dengan benar
3. Analisis Sumber data akuisisi berbagai jenis data dari berbagai sumber dengan benar
4. Pengumpulan data akuisisi berbagai jenis data dari berbagai sumber dengan benar

A. Defenisi dan Ruang Lingkup

Secara umum akuisisi data dapat diartikan sebagai proses atau kegiatan pengumpulan atau pemerolehan data. Defenisi tersebut sangat luas dan dapat menimbulkan perbedaan interpretasi. Untuk itu, dalam konteks Analis Data Ilmiah, akusisi data didefenisikan sebagai “proses atau kegiatan pengamatan fenomena alam, sosial, ataupun teknik secara langsung maupun tidak langsung dengan menggunakan peralatan atau alat bantu tertentu sehingga dihasilkan data yang dapat diolah oleh komputer untuk menjawab pertanyaan tertentu”. Berdasarkan defenisi ini, maka akusisi data dalam Analis Data Ilmiah mempunyai ciri sebagai berikut:

1. Merupakan proses atau kegiatan pengamatan fenomena alam, sosial, ataupun teknis. Dengan demikian pemerolehan data yang tidak melibatkan adanya kegiatan pengamatan fenomena tidak masuk dalam kategori ini. Misalnya, pemerolehan data melalui download dari Internet, pencarian dari basis data yang sudah ada, kesemuanya itu tidak termasuk dalam kegiatan akuisisi data.
2. Pengamatan dilakukan secara langsung maupun tidak langsung dengan

menggunakan peralatan atau alat bantu akuisisi tertentu. Penggunaan alat bantu menjadi penting untuk memastikan bahwa kegiatan akusisi data dilakukan secara benar. Misalnya, untuk pengukuran temperatur, jika hanya digunakan alat bantu kulit manusia maka data yang dihasilkan pasti berbentuk kualitatif dan bukan kuantitatif. Alat bantu akuisisi data juga menggambarkan proses pengamatan apakah secara langsung atau tidak langsung. Pengamatan langsung misalnya pada fenomena sosial dapatr menggunakan kuisioner, atau pengamatan fenomena alam menggunakan sendor. Pengamatan tidak langsung pada fenomena sosial misalnya melalui postingan individu di sosial media.

3. Data yang dihasilkan harus dapat diolah oleh komputer. Dalam hal ini data pada kuisioner yang masih berbentuk lembaran kertas belum termasuk data yang dapat dioleh oleh komputer.
4. Data dikumpulkan untuk menjawab pertanyaan tertentu. Ciri ini penting sebab data tidak bisa dikumpulkan secara acak atau tidak sistematis. Hal ini juga berlaku jika tiba-tiba menemukan data yang menarik di Internet tetapi tidak terkait dengan tugas dan fungsinya maka kegiatan tersebut tidak dapat dianggap sebagai kegiatan akuisisi data untuk Analis Data Ilmiah. Tujuan pengumpulan data sangat penting untuk memastikan bahwa data yang dikumpulkan sudah memenuhi kaidah ilmiah. Misalnya, jika tujuannya adalah untuk mengamati pengaruh bidang keilmuan terhadap tingkat sitasi karya tulis ilmiah, maka pengumpulan data harus memperhatikan keterwakilan dari setiap bidang keilmuan yang ingin diamati.

Untuk dapat diakui sebagai kegiatan akusisi data dalam Analis Data Ilmiah, maka ke-empat ciri tersebut harus ada dalam suatu kegiatan. Untuk mempermudah melihat perbandingan kegiatan akusisi data, beberapa contoh dan kesesuaiannya dapat dilihat pada Tabel 1.

Tabel 1. Contoh Kegiatan Akuisisi Data Yang Tidak Sesuai

No	Contoh Kegiatan	Ketidak Sesuaian
1	Memperoleh data dengan mendownload dataset dari repository data di Internet	Bukan merupakan hasil pengamatan fenomena alam, sosial, ataupun teknis
2	Memperoleh data dengan membeli dataset dari penyedia data	
3	Memperoleh data dari sharing dari kolega	
4	Memperoleh data dari Internet tetapi tidak mengetahui metodologi pengumpulannya	Data tidak dikumpulkan untuk menjawab pertanyaan tertentu
5	Memperoleh data secara acak dari Internet	

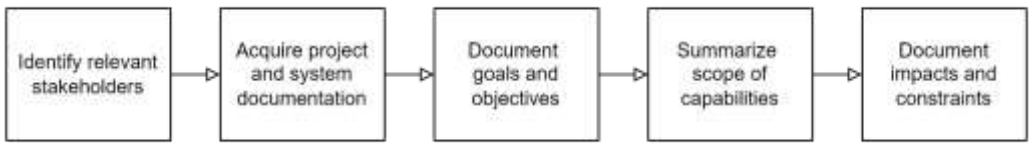
Untuk memenuhi ciri atau kriteria proses atau kegiatan akuisisi data tersebut, maka terdapat beberapa hal yang perlu diperhatikan terutama yang berkaitan dengan analisis kebutuhan/potensi data, analisis sumber data dan pengumpulan data.

B. Analisis kebutuhan / potensi data

Secara umum, analisis kebutuhan atau potensi data merupakan proses untuk mengidentifikasi, memprioritaskan, merumuskan secara tepat, dan memvalidasi data yang diperlukan untuk mencapai tujuan tertentu. Proses ini biasanya bersifat top-down yang digerakkan oleh kebutuhan, dimana kebutuhan tersebut perlu diidentifikasi dan pada saat yang sama perlu dipastikan kelayakannya yaitu kemungkinan kebutuhan tersebut akan dapat dipenuhi (Loshin, 2012). Proses analisis kebutuhan data terdiri dari beberapa tahapan:

1. Identifikasi konteks kebutuhan data

Konteks kebutuhan berhubungan dengan penentuan ruang lingkup hasil analisis yang akan dikontribusikan oleh data yang meliputi antara lain identifikasi data yang sudah ada atau penggunaan kembali data serta penentuan ukuran hasil pelaksanaan.



Gambar 1. Tahapan identifikasi konteks kebutuhan (diadopsi dari Loshin, 2012)

Gambar 1 menunjukkan tahapan identifikasi konteks kebutuhan yang umumnya dilakukan oleh seorang analis data yang mengerti akan kebutuhan data. Tahapan tersebut terdiri dari beberapa proses:

a. Identifikasi pihak yang terkait terutama pemakai data. Identifikasi dapat dilakukan melalui diskusi dengan expert yang mengerti tujuan kegiatan dan juga melalui dokumentasi kegiatan yang serupa.
b. Pendalaman dokumentasi untuk mengidentifikasi jenis-jenis sumber data termasuk apakah sumber data yang ada akan dapat memenuhi kebutuhan.
c. Identifikasi mekanisme yang digunakan untuk mengukur keberhasilan pencapaian target dan tujuan.
d. Rangkum ruang lingkup kebutuhan seperti fungsionalitas dan kapabilitas yang diinginkan.
e. Dokumentasikan dampak dan batasan-batasan yang dapat menghalangi tercapainya fungsionalitas dan kapabilitas yang ditargetkan.

2. Melakukan wawancara dengan stakeholder

Dokumentasi kegiatan yang serupa dapat menjadi sumber informasi untuk mengetahui ketersediaan data dalam institusi maupun diluar institusi. Informasi yang lebih banyak dapat diperoleh melalui pengguna yang akan menggunakan hasil analisis data tersebut. Tahapan ini dilakukan dengan melakukan diskusi dengan pengguna utama terutama yang akan menggunakan hasil analisis untuk mengambil keputusan.

Gambar 2. Tahapan wawancara dengan stakeholder (diadopsi dari Loshin, 2012)

Tahapan ini terdiri dari beberapa proses sebagaimana ditampilkan pada gambar 2, yaitu:

a. Identifikasi kandidat untuk diwawancarai dan peranan mereka. Proses ini penting untuk menentukan fokus diskusi yang sesuai dengan tanggung jawab masing-masing kandidat.

b. Menyusun pertanyaan wawancara. Membuat daftar pertanyaan yang dibutuhkan untuk memperoleh informasi tentang kebutuhan data. Ada 2 bentuk pertanyaan: (i) pertanyaan terarah untuk mengumpulkan informasi detail tentang fungsi dan proses dari suatu bagian kegiatan, (ii) pertanyaan terbuka untuk mengumpulkan informasi yang lebih umum dan bebas, biasanya dalam bentuk dialog.
c. Menyusun jadwal dan melakukan wawancara. Wawancara sebaiknya dilakukan tanpa adanya gangguan dari kegiatan yang lain.
d. Membuat kesimpulan dan mengidentifikasi kekosongan. Catatan dari wawancara dikumpulkan dan disusun sehingga dapat diketahui batasan-batasan kegiatan dan ketergantungan data dapat teridentifikasi.
e. Menyelesaikan kekosongan dan menyusun hasil akhir. Kesimpulan yang diperoleh dapat diklarifikasi kembali dengan para kandidat wawancara untuk memastikan kualitas data.

Hasil dari tahapan ini memungkinkan analis data untuk menyusun langkah-langkah maupun proses-proses yang dibutuhkan dalam suatu dokumen kebutuhan data.

3. Mengintegrasikan kebutuhan

Dokumen kebutuhan data yang diperoleh selanjutnya digunakan untuk mengidentifikasi konsep dan jenis data yang akan dikumpulkan dengan karakteristik data masing-masing.

Gambar 3. Tahapan integrasi kebutuhan (diadopsi dari Loshin, 2012)

Tahapan integrasi kebutuhan ditunjukkan pada gambar 3 dengan proses-proses yaitu:

a. Memodelkan aliran informasi. Model tersebut menggambarkan urutan, hirarki ataupun waktu proses yang terjadi dalam suatu kegiatan. Tujuannya adalah untuk memperoleh struktur maupun semantik untuk memastikan item data konsistent untuk konsolidasi ataupun aggregasi.

b. Mengindentifikasi kebutuhan elemen data. Kebutuhan konsep data seperti produk, perjanjian diidentifikasi melalui karakterisasi ataupun aggregasi kategori. Sebagai hasil diperoleh referensi data dan item master data potensial.
c. Spesifikasi kebutuhan fakta. Fakta merupakan bagian spesifik informasi yang akan dimonitor, dimanage, digunakan, dibagikan ataupun diteruskan dalam pelaksanaan kegiatan. Fakta tersebut juga dapat dilengkapi dengan kualifikasi atau dimensi data (misal waktu ataupun lokasi) yang dapat digunakan pada proses analisis selanjutnya.
d. Harmonisasi semantik elemen data. Dapat digunakan kumpulan metadata untuk menangkap terminologi yang digunakan. Kumpulan tersebut dapat disusun secara hirarki untuk mendukung komposisi ataupun aggregasu konsep data yang dianalisa. Kumpulan tersebut dapat digunakan untuk mengkonsolidasi dan harmonisasi termonilogi yang digunakan dalam institusi.

Sebagai hasil, data yang sudah diharmonsiasi dapat menghindari ketidakkonsistenan dalam pelaporan, analisis maupun kegiatan operasional sehingga dapat meningkatkan kepercayaan pada hasil analisis.

C. Analisis Sumber Data

Pada dasarnya, tahapan analisis sumber data bertujuan untuk memetaan sumber dan target data. Pada tahapan ini, ingin diidentifikasi 2 hal: (i) elemen data dari suatu sumber data yang berpotensi untuk diintegrasikan dalam proses analisis, (ii) mengetahui transformasi atau penyesuaian data yang diperlukan. Transformasi data menjadi penting pada tahapan ini untuk memastikan kesesuaian semantik data yang berasal dari berbagai sumber. Selain itu, keselarasan granularity data juga penting untuk memastikan proses analsisi data pada tahapan selanjutnya dilakukan secara benar.

Gambar 4. Tahapan pemetaan sumber dan target (diadopsi dari Loshin, 2012)

Tahapan pemetaan sumber dan target data ditampilkan pada gambar 4 yang terdiri dari:

a. Mengusulkan model pemetaan target data. Model ini merupakan data sharing model yang merepresentasikan elemen-elemen data yang akan diambil dari sumber-sumber yang tersedia, transformasi data yang dibutuhkan untuk validasi dan digunakan oleh aplikasi. Model tersebut dibangun dengan mempertimbangkan struktur konseptual ataupun logikal dari elemen-elemen data.
b. Mengidentifikasi kandidat sumber data. Mereview kandidat-kandidat sumber data yang mengandung elemen-elemen data serta fakta-fakta yang dibutuhkan untuk kegiatan analisis. Jika fakta merupakan hasil komputasi maka diidentifikasi elemen data yang digunakan untuk menghitung hasilnya tersebut. Setiap potential sumber data didokumentasikan.
c. Menentukan pemetaan sumber ke target data. Tahapan ini menentukan elemen data dari berbagai sumber ditransformasi dalam satu bentuk representasi yang seragam. Elemen data yang diperlukan dari setiap sumber data dikumpulkan termasuk transformasi yang dibutuhkan sesuai dengan standarisasi ataupun normalisasi yang teridentifikasi pada tahapan sebelumnya.

D. Pengumpulan Data

Terdapat berbagai metode untuk mengumpulkan data yang disesuaikan dengan sumber data maupun peralatan pengumpulan data. Metode tersebut diantaranya melakukan experimen, survei, interview atau fokus grup, observasi, traking, maupun monitoring sosial media. Berikut ini dijelaskan beberapa metode pengumpulan data disertai dengan peralatan yang umum digunakan.

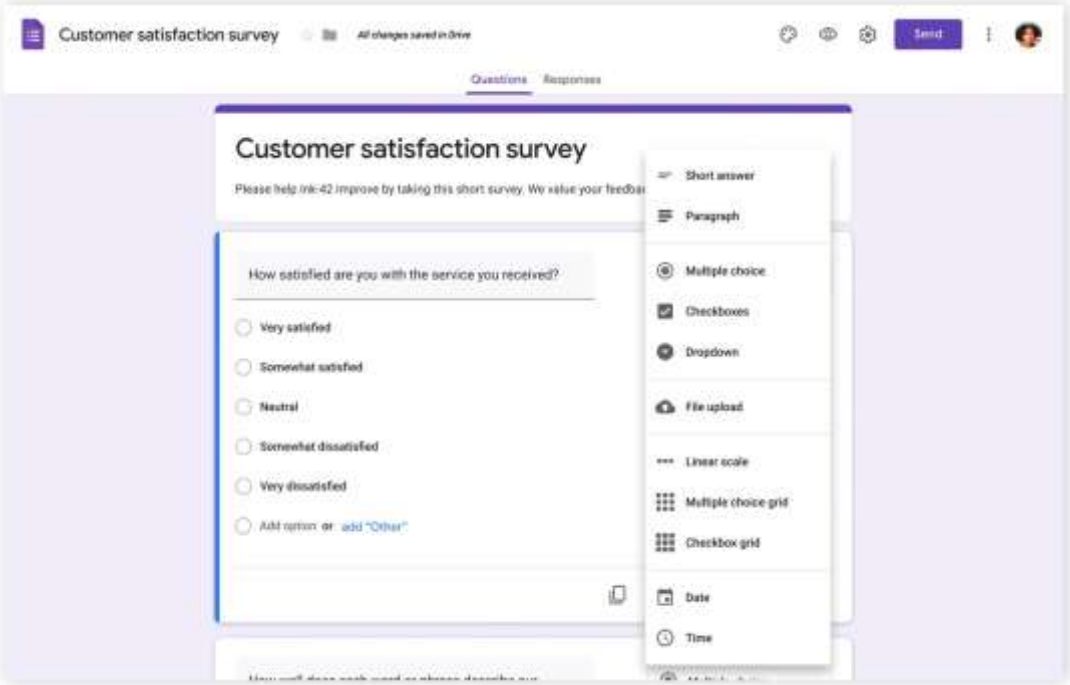
1. Melakukan eksperimen

Metode pengumpulan data yang paling dasar adalah melalui eksperimen yaitu pelaksaanan prosedur ilmiah untuk menemukan hal baru, mengetes hipotesa ataupun mendemonstrasikan fenomena yang diketahui. Pelaksanaan tersebut umumnya dilakukan pada lingkungan yang terkontrol seperti laboratorium. Peralatan pengumpulan data yang umum digunakan misalnya dari alat sederhana seperti mikroskop, stetoscope, atmometer,

sampai pada alat yang komplek seperti spektroskopi, mesin genome sequencing. Data yang dihasilkan dapat berbentuk numerik, tekstual, gambar, audio, ataupun video.

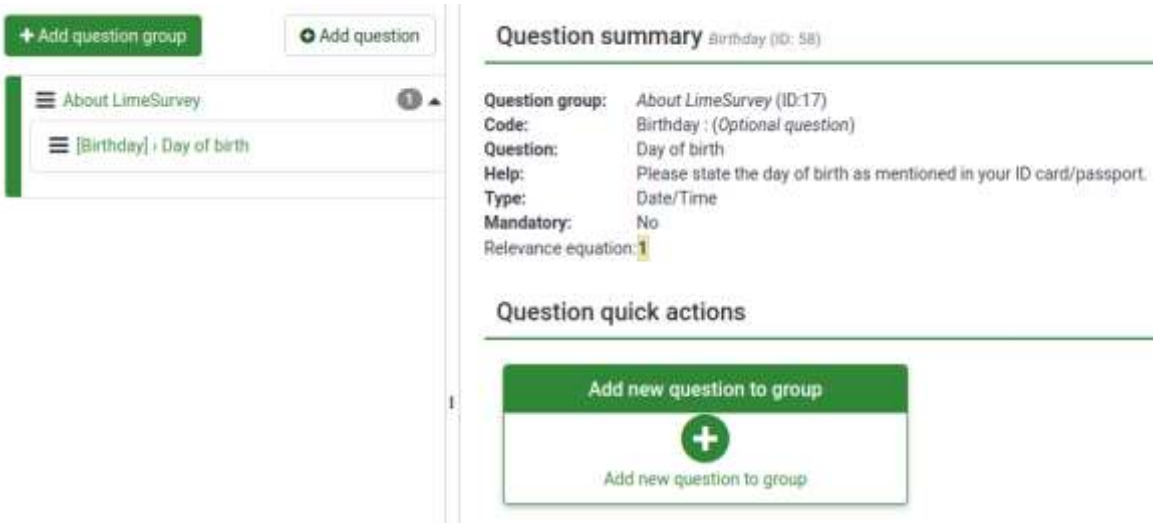
2. **Survei**

Pada dasarnya survei bertujuan untuk mengumpulkan informasi dari sekelompok orang tentang opini, kelakuan maupun pengetahuan mereka. Pelaksanaan survei dapat melalui kuisioner tertulis, interview melalui tatap muka langsung ataupun melalui telepon, fokus grup, ataupun melalui media elektonik seperti email atau website. Penggunaan peralatan survei melalui media elektronik terutama website semakin banyak digunakan karena kemudahan yang ditawarkan. Contohnya adalah Google Form[1], Survey Monkey[2], Lime Survey[3]. Gambar 5 menunjukkan tampilan dari Google Form dimana berbagai bentuk pertanyaan umum (misalnya pilihan ganda, checkbox, ataupun essay) dapat difasilitasi. Demikian juga dengan Lime Survey seperti yang ditampilkan pada gambar 6, tersedia berbagai bentuk pertanyaan bahkan aplikasi ini bisa diinstal pada server sendiri.



Gambar 5. Contoh halaman Google Form (sumber: Google.com)

[1] https://www.google.com/forms/about/
[2] https://www.surveymonkey.com/
[3] https://www.limesurvey.org/

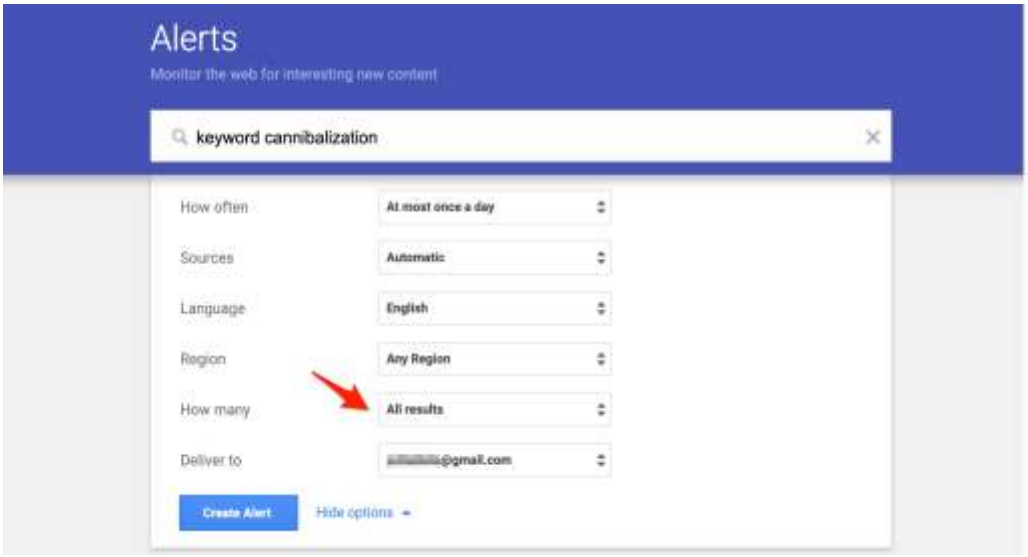


Gambar 6. Contoh halaman Lime Survey (sumber: Limesurvey.org)

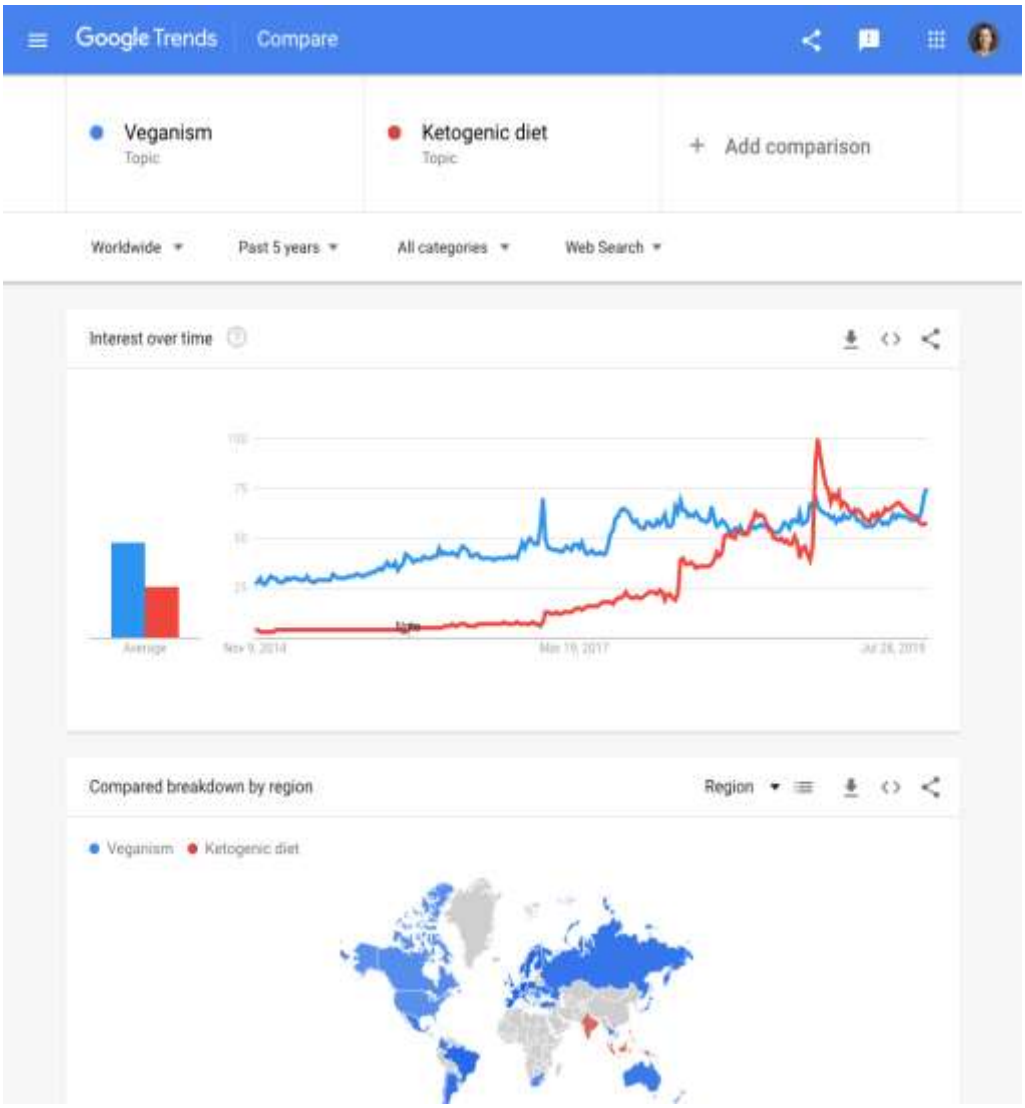
3. Web monitoring

Web telah menjadi sumber data yang sangat besar dan beragam. Pengumpulan data melalui web membutuhkan metode yang sistematik untuk menemukan infomasi yang relevan. Karena volume data yang tersedia di Internet sudah sangat besar, berbagai alat bantu telah disediakan terutama oleh perusahaan mesin pencari, misalnya Google Alerts [4] dan Google Trends[5]. Google Alerts dapat memberikan layanan berupa pemberitahuan melalui email tentang perubahan atau informasi baru yang tersedia di Internet yang sesuai dengan kata kunci yang diberikan oleh pengguna. Sementara itu, Google Trends dapat memberikan informasi tentang topik pencarian yang populer pada berbagai region dan bahasa pada rentang waktu tertentu. Gambar 7 menunjukkan antar muka untuk Google Alerts sedangkan tampilan Google Trends ditampilkan pada gambar 8.

[4] https://www.google.com/alerts
[5] https://trends.google.com/



Gambar 7. Contoh halaman Google Alerts (sumber: google.com)

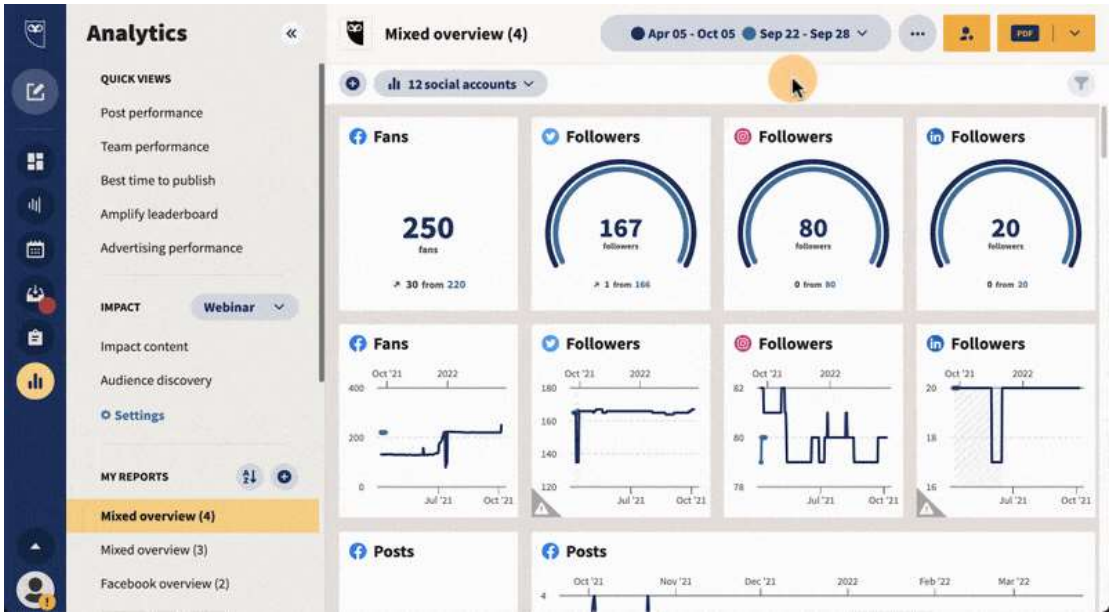


Gambar 8. Contoh halaman Google Trends (sumber: google.com)

4. **Sosial media monitoring**

Sosial media monitoring adalah salah satu metode pengamatan fenoma sosial secara tidak langsung dengan mengumpulkan data melalui informasi yang ada pada sosial media. Data yang dapat dikumpulkan tergantung pada jenis media yang difasilitasi termasuk fitur-fitur interaksi sosial yang disediakan oleh setiap platform sosial media. Jenis media data misalnya tekstual, gambar, audio, video termasuk dengan metadatanya seperti geolokasi, waktu pengambilan data, kamera yang digunakan untuk mengambil gambar, dll. Selain itu, fitur-fitur sosial seperti like, dislike, follow, comment, juga menjadi sumber data yang kaya.

Pengumpulan data terkait aktifitas di berbagai platform sosial media dapat dilakukan secara otomatis dengan menggunakan alat bantu, baik yang disediakan oleh platform itu sendiri maupun pihak ketiga. Gambar 9 menunjukkan contoh alat bantu untuk memonitor aktifitas sosial media diberbagi akun dan platform. Berbagai jenis data dapat diakusisi sesuai dengan fasilitas yang didukung oleh masing-masing platform.

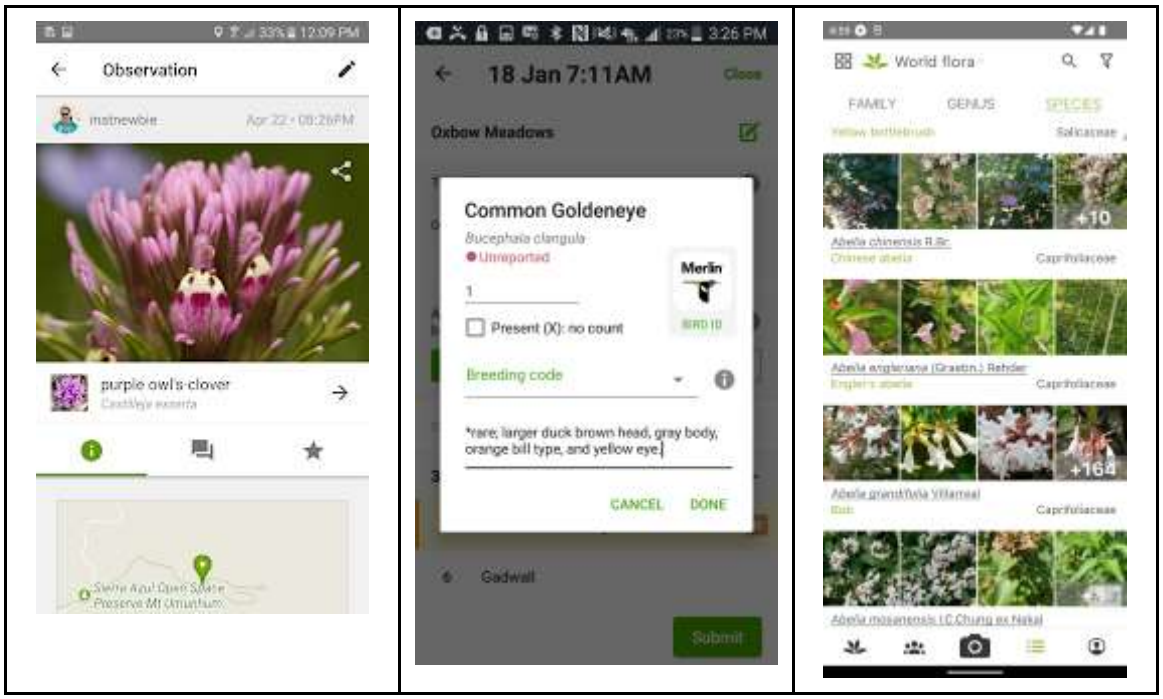


Gambar 9. Contoh halaman Hootsuite Analytics (sumber: hootsuite.com)

5. **Crowdsourcing**

Secara umum, crowdsourcing adalah mekanisme untuk memperoleh ide, data, informasi, maupun layanan dengan meminta bantuan dari orang lain secara

massal. Kontribusi dari partisipan biasanya dilakukan secara sukarela. Salah satu bentuk crowdsourcing adalah Citizen Science[6], dimana masyarakat umum dilibatkan dalam kegiatan explorasi ilmu pengetahuan termasuk diantaranya pengumpulan data (Gura, 2013). Pengumpulan data melalui mekanisme ini telah banyak digunakan diberbagai dibidang ilmu terutama lingkungan dan ekologi (Fraisl et al., 2022). Mekanisme pengumpulan data umumnya dilakukan dengan menggunakan alat bantu terutama smartphone. Gambar 10 menampilkan contoh beberapa aplikasi smartphone untuk pengumpulan data berbasis citizen science: Inaturalist[7] (kiri) untuk pengamatan flora dan fauna secara umum, eBird[8] (tengah) untuk pengamatan khusus burung dan Pl@ntnet[9] untuk identifikasi dan observasi tumbuhan (kanan).



Gambar 10. Contoh aplikasi smartphone untuk mengumpulkan data dari masyarakat (sumber: inaturalist.org, ebird.org, plantnet.org)

Selain itu, aplikasi pengumpulan data berbasis citizen science juga dapat berupa kombinasi antara perangkat lunak dan keras. Hal itu dibutuhkan

[6] https://en.wikipedia.org/wiki/Citizen_science
[7] https://www.inaturalist.org/
[8] https://ebird.org/
[9] https://identify.plantnet.org/

terutama jika kemampuan dari smartphone tidak mencukupi sehingga dibutuhkan tambahan perangkat keras lainnya. Sebagai contoh, gambar 11 menunjukkan aplikasi yang disebut HABscope (Hardison et al., 2019) yaitu mikroskop digital portabel yang digunakan untuk mengamati jenis dan kuantitas alga dalam air untuk mengukur potensi kejadian Harmful Algal Blooming (HAB)[10].

Gambar 11. Contoh alat pengumpulan data lapangan, kombinasi antara perangkat lunak dan perangkat keras (sumber: Hardison et al., 2019)

[10] https://en.wikipedia.org/wiki/Harmful_algal_bloom

E. Rangkuman

Akuisisi data merupakan tahapan penting dalam kegiatan Analis Data Ilmiah. Data yang dihasilkan akan digunakan dalam tahapan-tahapan selanjutnya sehingga kualitas data yang diakuisis akan menentukan kualitas hasil analisa nantinya. Akuisisi data tidak semata-mata terkait pengumpulan data semata, ada kegiatan lain yang perlu dilakukan terlebih dahulu yaitu analisa kebutuhan/potensi data serta analisa sumber data. Analisa kebutuhan/potensi data perlu dilakukan untuk mengukur ketersediaan data baik internal maupun eksternal. Sedangkan analisa sumber data dilakukan untuk mengatahui sumber data yang paling sesuai dengan kebutuhan. Untuk pengumpulan data itu sendiri, dapat dilakukan secara langsung maupun tidak langsung melalui alat bantu yang tersedia. Pengembangan alat bantu ini juga penting untuk memastikan data yang terkumpul sesuai dapat menjawab pertanyaan kegiatan analis data ilmiah yang sedang dilakukan.

F. Evaluasi

1. Identifikasi salah satu kegiatan terkait analis data ilmiah pada unit kerja atau instansi anda?
2. Dari kegiatan yang teridentifikasi pada nomor 1, lakukan:
   a. Analisa kebutuhan/potensi anda
   b. Analisa sumber data
3. Dari hasil analisa kebutuhan/potensi serta sumber data dari nomor 2, tentukan alat bantu pengumpulan data yang sesuai, salah satunya gunakan alat bantu survei.
4. Buat survei menggunakan tool yang sesuai untuk nomor 3!

MATERI POKOK 2:

PELABELAN DATA

Indikator Hasil Belajar:

Peserta mampu **menjelaskan pentingnya label data untuk pengolahan data secara otomatis dengan benar**, meliputi:

1. Menjelaskan definisi dan ruang lingkup label data untuk pengolahan data secara otomatis dengan benar.
2. Menjelaskan metode pelabelan data untuk pengolahan data secara otomatis dengan benar

A. Defenisi dan Ruang Lingkup

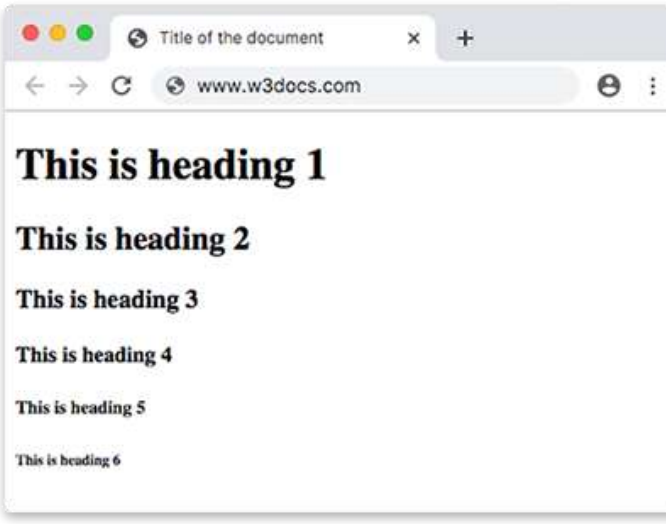
Pelabelan data atau dikenal juga sebagai data tagging atau data annotation dapat diartikan sebagai proses pemberian label yang lebih bermakna dan informatif pada data mentah. Dalam kegiatan Analis Data Ilmiah, pelabelan data dimaksudkan untuk memberikan konteks pada data mentah sehingga data tersebut dapat dimengerti dan diolah dengan benar baik oleh manusia maupun mesin. Selain itu label tersebut juga memberikan pemahaman yang seragam pada elemen maupun item data sehingga menghindari ketidakkonsistenan dalam kegiatan analisis nantinya. Berdasarkan uraian ini, maka pelabelan data mempunyai ciri sebagai berikut:

- Memberikan konteks pada data mentah. Dalam hal ini konteks merupakan informasi latar belakang yang memberikan pemahaman yang lebih luas pada data. Misalnya informasi tentang waktu yang memberikan pemahaman kapan data dibuat atau kapan diubah, informasi tentang penulis yang memberikan pemahaman siapa yang membuat atau siapa yang merubah, dll.
- Tujuan pelabelan adalah supaya data dapat dimengerti dan diolah dengan benar oleh manusia dan mesin. Label data akan memberikan informasi tambahan yang memperkaya informasi yang sudah ada. Label tersebut digunakan baik oleh manusia maupun mesin dalam pengolahannya. Hal tersebut juga memungkinkan data untuk dipahami secara seragam, mencegah terjadinya ambigu yang dapat mempengaruhi kualitas hasil

pengolahan data.

Secara umum, hasil pelabelan data adalah berupa informasi tambahan yang terpisah dari data mentah, dimana satu data mentah dapat mempunyai 1 atau lebih label (multi-label) sesuai dengan konteks yang ingin diberikan. Pelabelan data ini menjadi sangat penting terutama pada lingkungan dimana data bersifat heterogen dan terdistribusi sebagai terjadi pada web. Sebagai contoh, adalah pemberian label (tag) headings yang berbeda pada halaman web sehingga data akan divisualisasikan secara berbeda seperti terlihat pada gambar 12. Dengan memberikan tag “<h1>”, maka text akan divisualisasikan menjadi lebih besar dengan tag “<h2>” samapi “<h6>”.

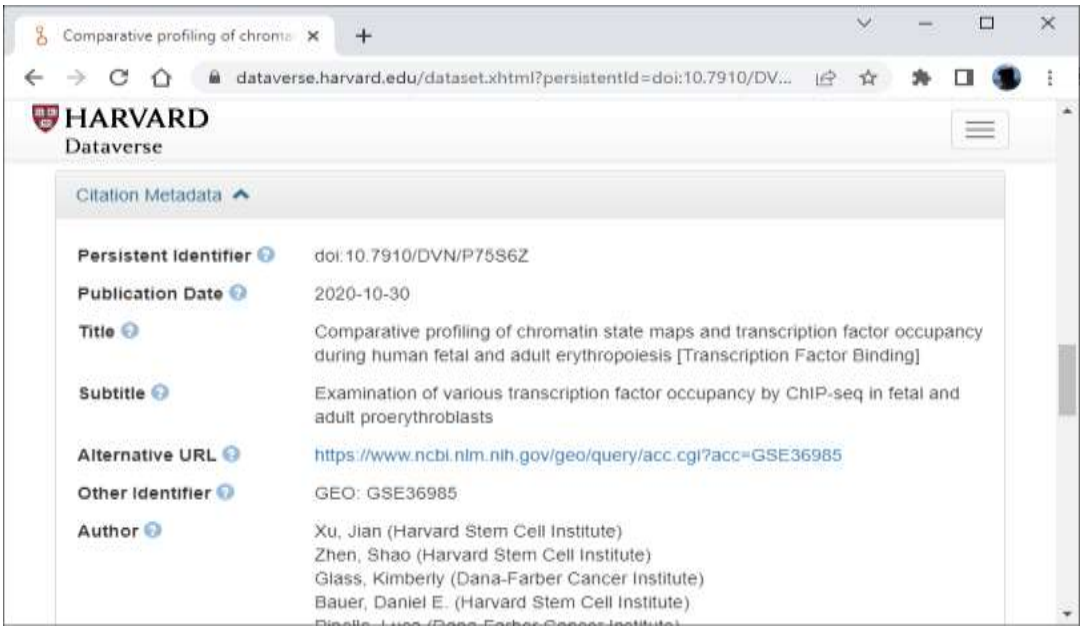
```
<h1>This is heading 1</h1>
<h2>This is heading 2</h2>
<h3>This is heading 3</h3>
<h4>This is heading 4</h4>
<h5>This is heading 5</h5>
<h6>This is heading 6</h6>
```



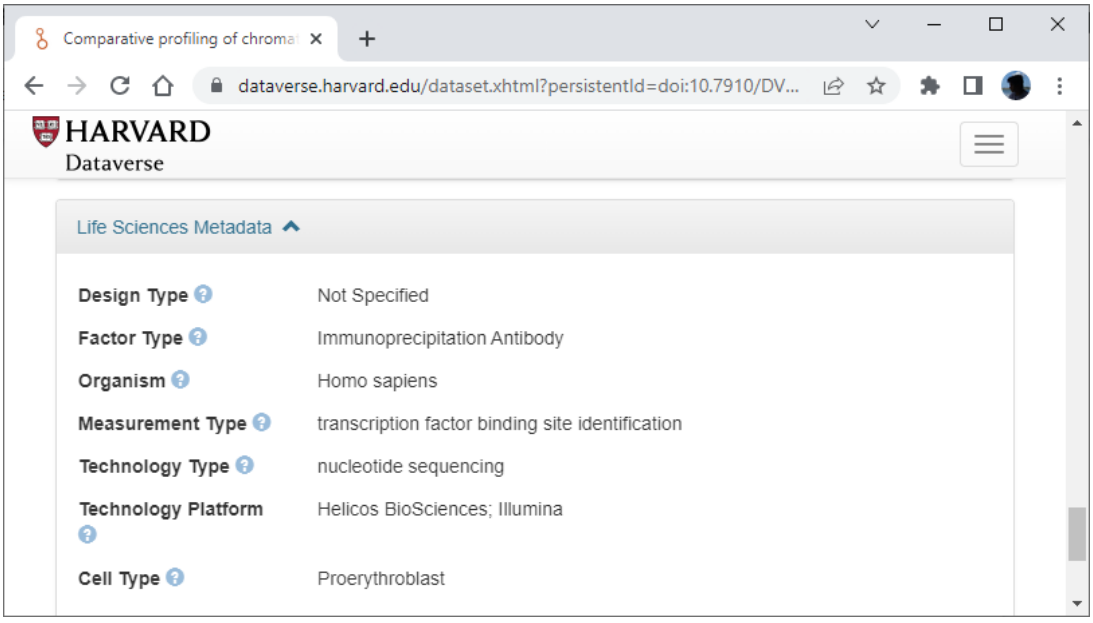
Gambar 12. Contoh pemberian label untuk tujuan visualisasi

Pelabelan data juga dapat dimaksudkan untuk memberikan informasi tambahan terkait proses produksi dari suatu data sehingga dapat diakui atau direferensi dengan benar oleh pemakai. Gambar 13 memberikan contoh metadata yang dapat digunakan untuk sitasi, diantaranya memuat informasi judul atau topik dataset, tanggal di publikasi, penulis, tanggal dan waktu dataset dideposit, publikasi atau dataset lain yang terkait, dst. Dataset yang sama dapat diberi label yang berbeda dengan tujuan yang berbeda pula. Sebagai contoh, pada gambar 14 menampilkan metadata untuk dataset yang sama dengan yang ditampilkan pada gambar 13. Namun demikian, metadata pada gambar 14 lebih ditujukan untuk pengarsipan data sehingga memudahkan dalam pencarian data nantinya.

Dalam contoh ini. metadata yang digunakan terkait dengan ilmu hayati seperti jenis organisme yang diukur, jenis pengukuran, teknologi dan platform pengukuran yang digunakan, dll.

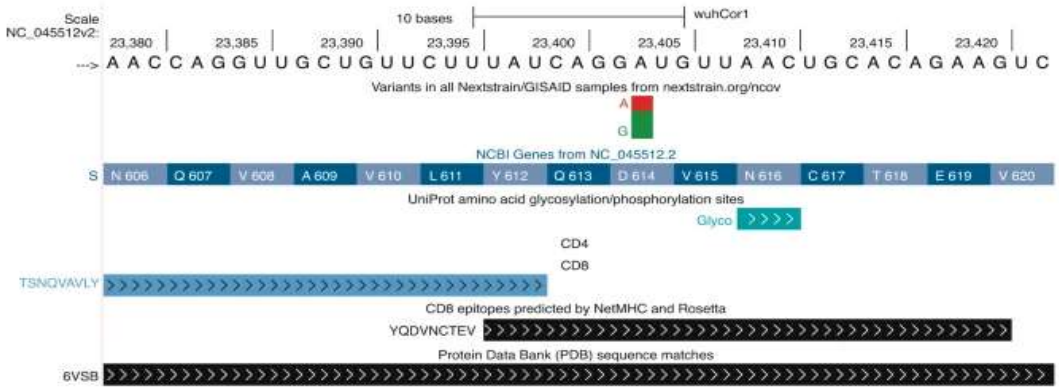


Gambar 13. Contoh pemberian label untuk tujuan sitasi (sumber: dataverse.harvard.edu)



Gambar 14. Contoh pemberian label untuk tujuan pengarsipan data sehingga memudahkan pencarian (sumber: dataverse.harvard.edu)

Dalam kasus analis data, pelabelan data dimaksudkan lebih dari tujuan visualisasi seperti contoh sebelumnya. Tujuan utamanya adalah bagaimana supaya pengolahan data yang dilakukan dapat memberikan hasil yang benar dan lebih baik. Tantangan pengolahan data misalnya ukuran data yang sangat besar misalnya urutan genome seperti yang ditampilkan pada gambar 15.

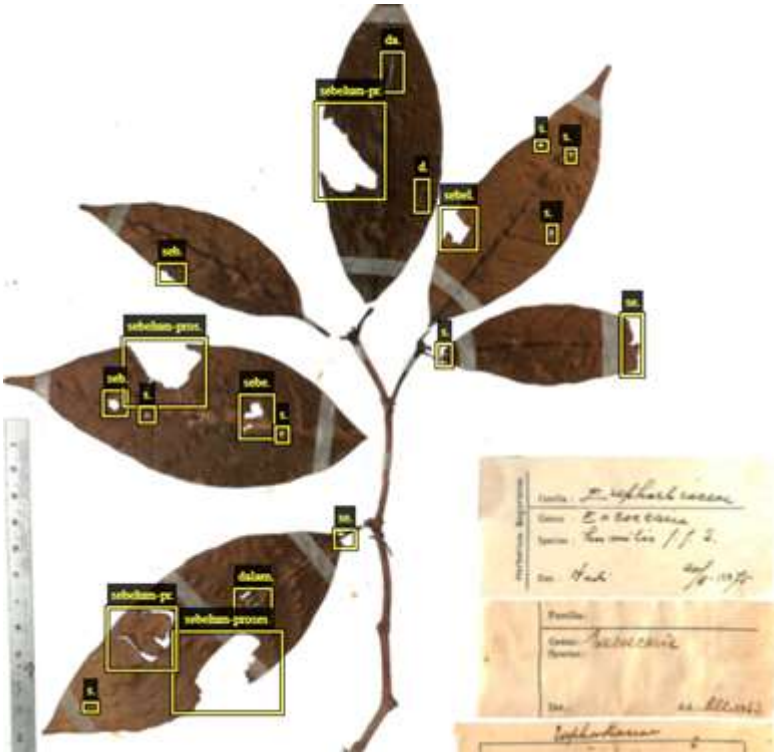


Gambar 15. Contoh annotation untuk urutan genome (sumber: Fernandes et al., 2020)

Mengingat adanya berbagai macam maksud dan tujuan pelabelan data, maka pada materi ini hal tersebut dibatasi pada tujuan analis data dimana label diberikan sebagai informasi tambahan pada data. Pelabelan data jenis ini umumnya ditemukan pada kasus pembelajaran mesin secara otomatis yang menggunakan teknik *supervised learning*.

1. **Pelabelan Data: Identifikasi Obyek**

   Tujuan pelabelan data yang umum dijumpai adalah untuk identifikasi obyek. Dalam hal ini obyek-obyek yang menjadi perhatian dalam data akan diberi label yang sesuai sehingga mesin dapat menggunakan label tersebut sebagai referensi. Gambar 16 menunjukkan contoh annotasi menggunakan pendekatan *region-of-interest*, dimana obyek yang menjadi perhatian (dalam hal ini kerusakan akibat serangga) aka diberi tanda berupa *bounding-box* (menentukan koordinat serta luasan) yang disertai dengan labelnya (dalam hal ini jenis-jenis kerusakan). Hal serupa juga ditunjukkan pada gambar 17, tetapi pada contoh ini diannotasi berbagai jenis obyek seperti jendela, pintu, mobil, pohon, rumah, manusia, sehingga menghasilkan multi-label.

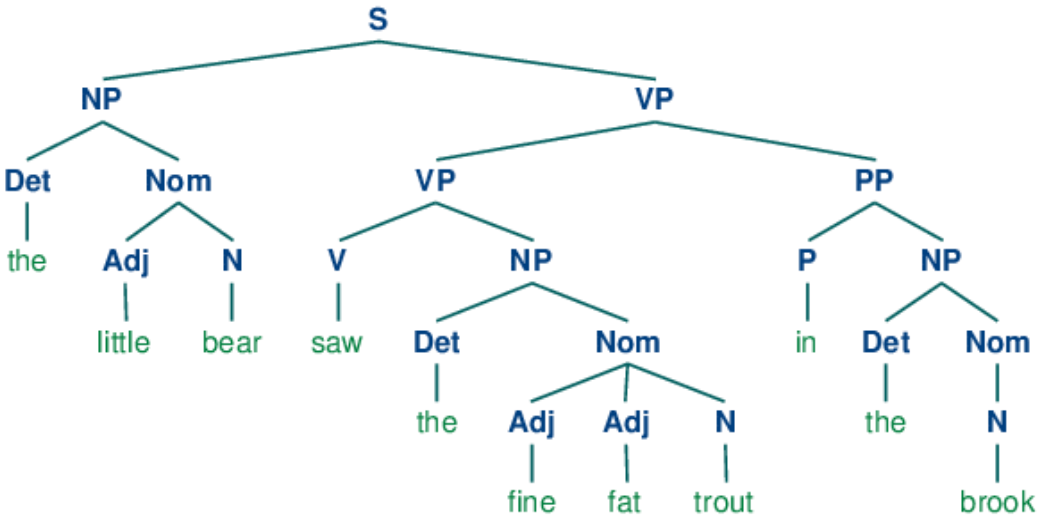


Gambar 16. Contoh annotasi gambar spesimen tanaman untuk identifikasi obyek kerusakan yang diakibatkan oleh serangga (sumber: Akbar et al., 2022)

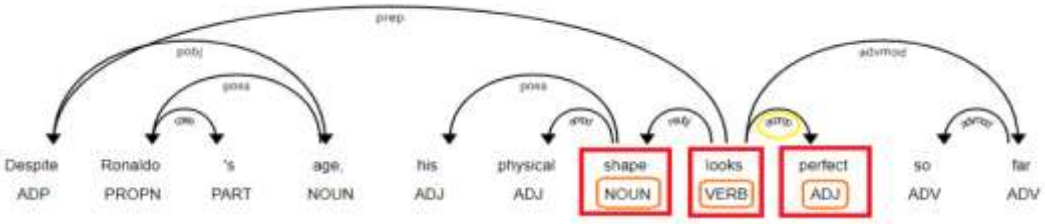
Gambar 17. Contoh annotasi gambar pemandangan luar ruangan untuk identifikasi berbagai jenis obyek (sumber: Russell et al., 2008)

2. **Pelabelan Data: Identifikasi Relasi dan Struktur**

Pelabelan data dapat juga digunakan untuk identifikasi relasi antar entitas maupun struktur entitas. Contoh yang umum ditemukan adalah pada pengolahan data bahasa alami secara otomatis. Dalam hal ini kata-kata dalam suatu kalimat akan diberi label sesuai dengan arti dan juga konteksnya. Misalnya pelabelan *part-of-speech* (PoS) akan memberikan label gramatik tertentu (kata benda (*noun*), kata ganti (*pronoun*), kata kerja (*verb*), kata sifat (*adjective*), kata keterangan (*adverb*), kata depan (*preposition*), kata hubung (*conjunction*), dan kata seru (*interjection*)) pada setiap kata sehingga pohon kalimat dapat dibangun. Pohon kalimat tersebut dapat digunakan untuk berbagai hal, misalnya untuk membuat summarisasi dari tulisan, sistem tanya jawab.



Gambar 18. Contoh pelabelan untuk memperoleh struktur kalimat (sumber: Bird et al., 2009)



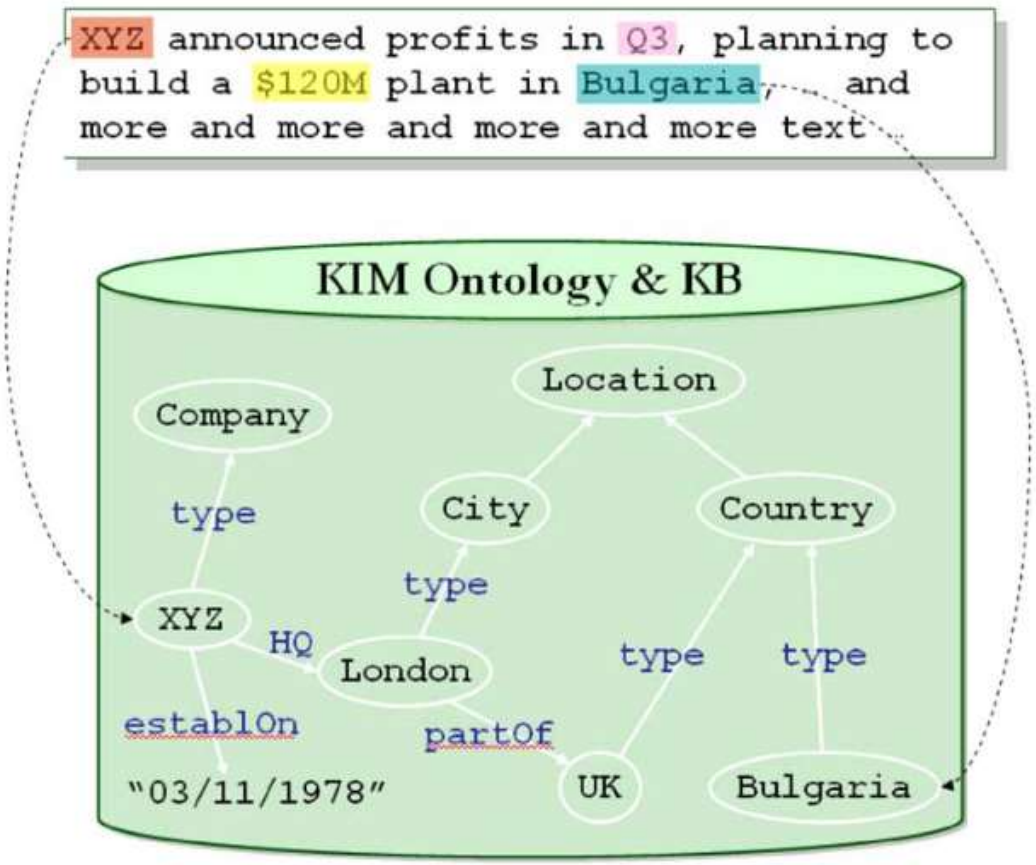
Gambar 19. Contoh annotasi bagian dari kalimat (sumber: Consoli et al., 2021)

Gambar 18 menunjukkan contoh pelabelan kata untuk memperoleh struktur kalimat dimana struktur kalimat tersebut selanjutnya dapat digunakan untuk proses analisa data selanjutnya. Sebagai contoh, struktur tersebut digunakan untuk melakukan analisa sentimen seperti yang ditampilkan pada gambar 19.

3. **Pelabelan Data: Identifikasi Semantik**

   Semantik mengacu pada ilmu yang mempelajari arti atau makna dari kata ataupun kalimat, dimana umumnya makna akan terkandung dari hubungan antar kata atau kalimat tersebut. Hal tersebut penting terutama dalam pengolahan bahasa alami dimana ada kata yang mempunyai makna yang berbeda atau ada beberapa kata yang mempunyai makna yang sama. Sebagai contoh, kata “Jaguar” dapat merepresentasikan dua hal yang berbeda, binatang[11] atau mobil[12]. Untuk menghindari ambiguitas tersebut dapat digunakan label untuk menghilangkan hal tersebut. Gambar 20 menunjukkan contoh pemberian label untuk meng-identifikasi entitas secara secara benar. Dalam hal ini digunakan Ontology[13] dan basis pengetahuan (*Knowledge Base* (KB)[14]) sebagai referensi. Pertama-tama, XYZ itu akan diidentifikasi sebagai perusahaan (*Company*) dan Bulgaria itu sebagai negara (*Country*). Berdasarkan KB yang dibangun diketahui bahwa XYZ itu mempunyai kantor pusat (HQ) di London yang merupakan kota (*City*) sebagai bagian dari (*partOf*) dari UK yang merupakan negara (*Country*). Dengan demikian maka tidak akan ada lagi ambiguitas lagi terkait dengan XYZ. Contoh penggunaanya untuk pencarian presisi, dimana jika ada yang mencari XYZ yang merupakan nama orang, maka XYZ dalam contoh ini tidak termasuk dalam hasil pencarian.

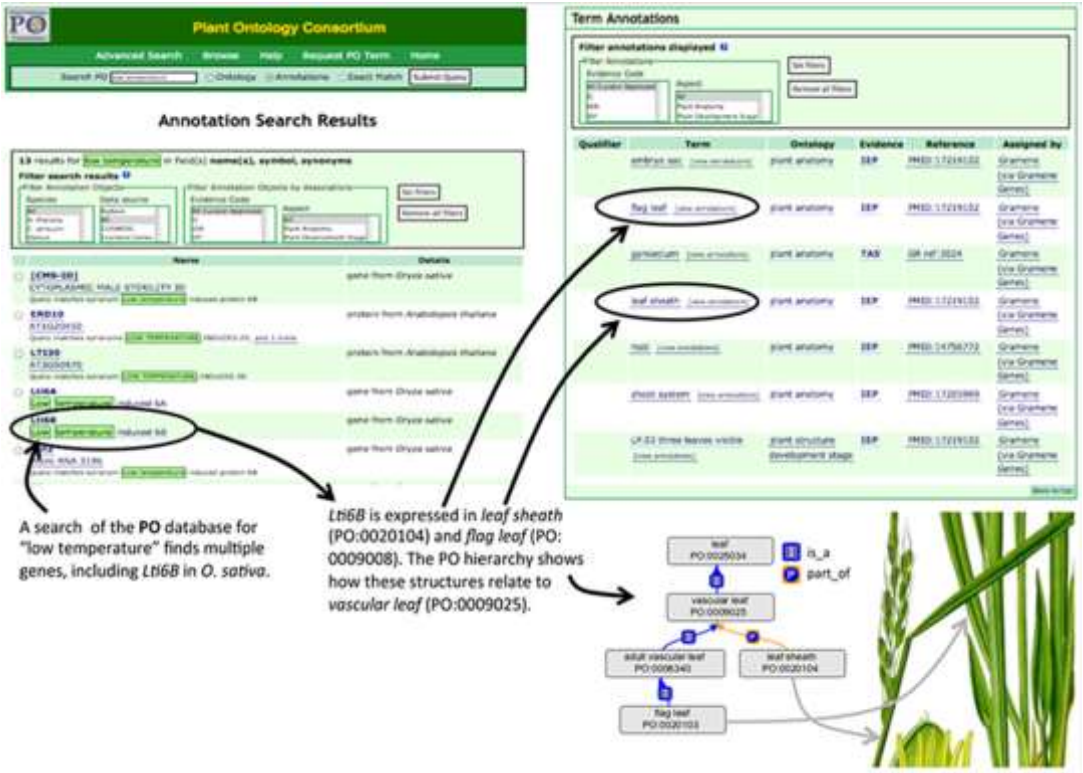
[11] https://en.wikipedia.org/wiki/Jaguar
[12] https://en.wikipedia.org/wiki/Jaguar_Cars
[13] https://en.wikipedia.org/wiki/Ontology_(information_science)
[14] https://en.wikipedia.org/wiki/Knowledge_base



Gambar 20. Contoh annotasi semantik (sumber: Kiryakov et al., 2004)

Contoh lainnya ditunjukkan pada gambar 21, dimana teks yang berisi anatomi tanaman akan diberi label menggunakan Plant Ontologi (PO) (Jaiswal et al., 2005). Misalnya sarung daun (*leaf sheath*) itu dilabeli PO:0020104[15], bendera daun (*flag leaf*) dengan PO:0009008[16]. Sebagai hasilnya, tidak akan terjadi lagi ambiguitas dalam penggunaan terminologi tersebut karena mengacu pada referensi yang sudah distandarisasi. Lebih dari itu, penggunaan ontologi ini memungkinkan dilakukannya deduksi kesimpulan, misalnya berdasarkan label tersebut, dalam PO diketahui bahwa struktur tanaman tersebut menunjuk ke daun vaskular (*vascular leaf*) yang dapat dilihat berdasarkan strukturnya (PO:0009025[17]).

[15] https://browser.planteome.org/amigo/term/PO:0020104
[16] https://browser.planteome.org/amigo/term/PO:0009008
[17] https://browser.planteome.org/amigo/term/PO:0009025



Gambar 21. Annotasi teks menggunakan terminologi dari ontology (sumber: Walls et al., 2012)

B. Metode Pelabelan Data

Dibagian sebelumnya telah dijelaskan beberapa tujuan pelabelan data, diantaranya untuk memungkinkan dilakukannya identifikasi obyek-obyek tertentu, identifikasi relasi antara entitas maupun struktur, identifikasi arti kata maupun arti relasi (semantik) yang terkandung dalam data. Metode untuk melakukan pelabelan tersebut dapat dilakukan secara manual maupun otomatis.

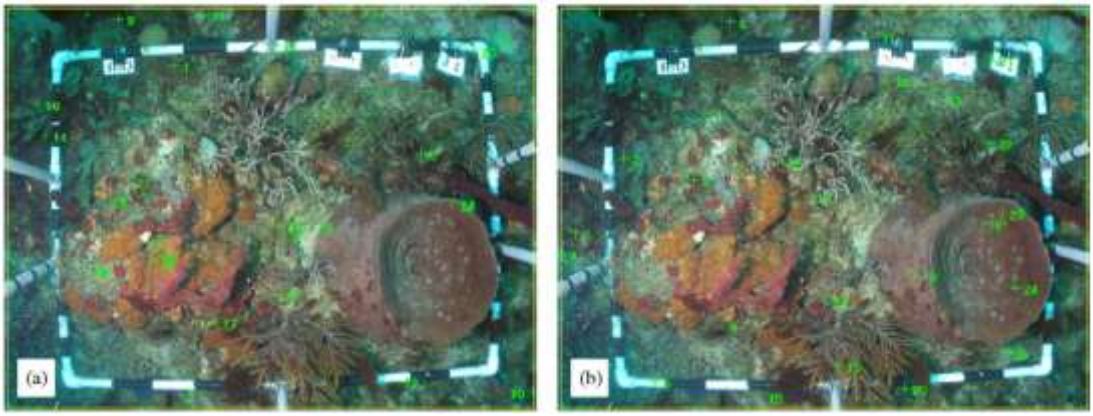
1. Pelabelan secara manual

Pelabelan ini umumnya menggunakan ekspert yang memberikan label tertentu dengan menggunakan alat bantu. Dalam beberapa kasus, label akan diberikan untuk keseluruhan item data, misalnya apakah tweet ini mempunyai sentimen positif, negatif, atau netral. Namun jika pelabelan dilakukan pada obyek-obyek tertentu dari data, maka expert akan menandai obyek tersebut lalu memberikan label. Gambar 22 menunjukkan contoh pelabelan untuk obyek dalam multimedia (gambar, audio, dan video). Untuk gambar, obyek yang diinginkan (swan) akan ditandai lalu diberi keterangan. Untuk audio, data akan dipecah sesuai dengan waktu dan berikan label, misalnya siapa

yang berbicara antara waktu tertentu. Sedangkan untuk video merupakan gabungan antara pelabelan gambar dan audio.

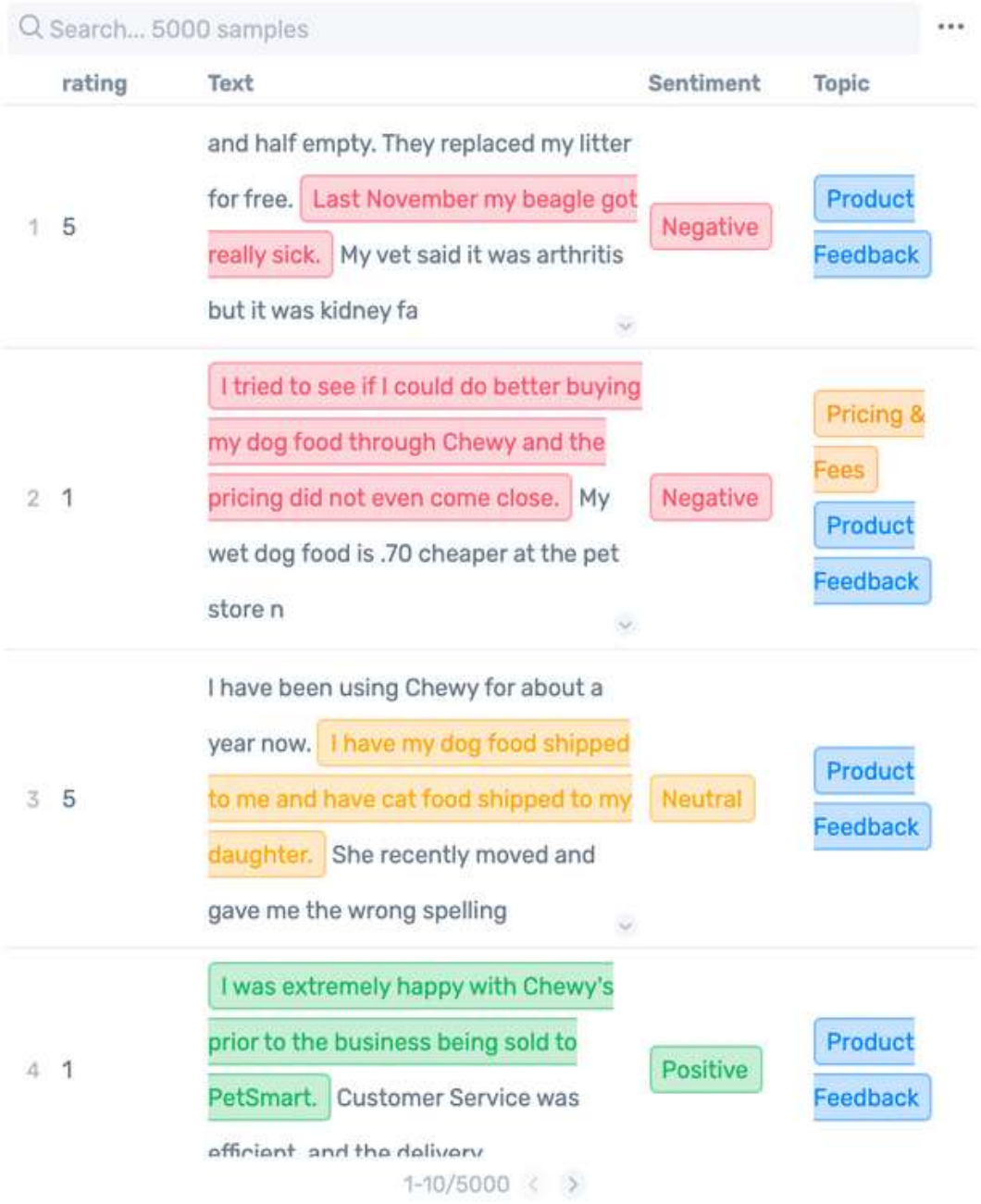
Gambar 22. Pelabelan data secara manual menggunakan VIA Annotation Software untuk melabeli gambar (kiri), audio (tengah), video (kanan) (sumber: Dutta & Zisserman, 2019)

Terdapat berbagai macam teknik pelabelan selain yang disampaikan sebelumnya. Gambar 23 menunjukkan contoh pelabelan koral dengan membuat titik-titik tertentu pada gambar dan memberikan keterangan untuk setiap titiknya. Berdasarkan intensitas cahaya dari setiap titik, komputer akan mencari piksel yang serupa untuk diberi label yang sama.



Gambar 23. Pelabelan data secara manual menggunakan software Coral Point Count with Excel extensions (CPCe) untuk identifikasi koral (sumber: Kohler & Gill, 2006)

Contoh lainnya adalah untuk pelabelan sentimen dari teks seperti review, komentar, ataupun postingan sosial media. Gambar 24 menunjukkan contoh pelabelan tersebut, dimana teks akan diberikan sentimen positif, negatif, maupun netral yang sesuai.

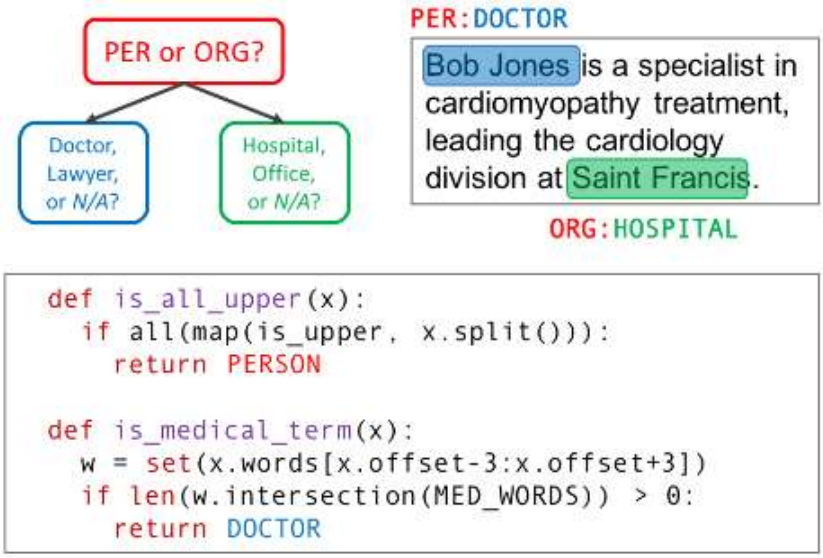


Gambar 24. Contoh pelabelan sentimen (positif, negatif, netral) untuk teks (sumber: monkeylearn.com)

2. **Pelabelan secara otomatis**

   Metode pelabelan kedua adalah metode otomatis dimana label diberikan oleh expert secara tidak langsung melalui teknik tertentu. Salah satu teknik yang dapat digunakan disebut *Weak Supervision* (Ratner et al., 2020). Pada teknik ini, expert akan mendefiniskan berbagai fungsi pelabelan (seperti aturan pola

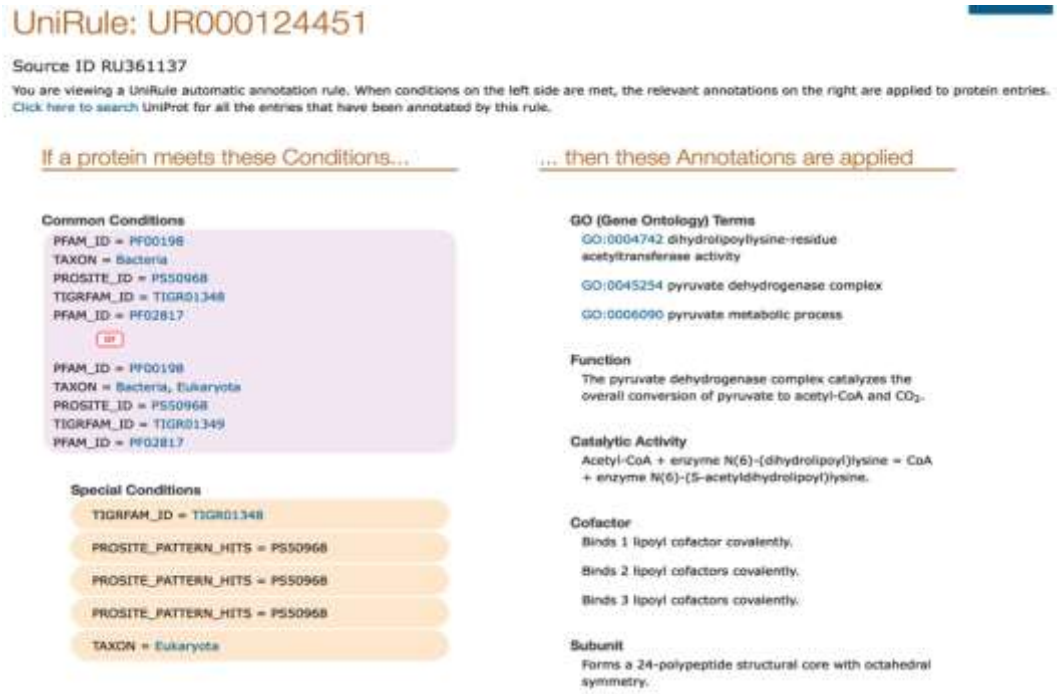
secara heuristik, pencarian kamus) untuk digunakan memberikan label. Label yang dihasilkan oleh seluruh fungsi dintegrasikan untuk menentukan label yang sebenarnya. Gambar 25 menunjukkan contoh pelabelan menggunakan teknik *Weak Supervision*, dimana dua fungsi pelabelan yang relatif sederhana (*weak*) digunakan untuk melabeli apakah kata (atau kumpulan kata) itu merepresentasikan entitas Orang atau Organisasi. Fungsi pelabelan pertama akan mengecek apakah semua kata itu huruf kapital, jika iya maka akan dilabeli sebagai PERSON. Fungsi pelabelan kedua akan mengecek kesesuain kata dalam kamus kedokteran, jika iya, maka orang tersebut adalah DOCTOR.



Gambar 25. Pelabelan data secara otomatis menggunakan teknik *Weak Supervision* (sumber: Ratner et al., 2019)

Teknik pelabelan otomatis lainnya adalah melalui *rule-based system*, yaitu sistem yang menggunakan sekumpulan aturan (*rules*) untuk menarik kesimpulan. Rules tersebut umumnya didefinisikan oleh expert dibidang tertentu, biasanya dalam format IF <kondisi> THEN <konsekuensi>. Rules tersebut selanjutnya dikonsumsi oleh mesin (*rules engine*) untuk melakukan penelusuran secara deduktif sampai diperoleh <konsekuensi> paling akhir sebagai hasil. Gambar 26 menunjukkan contoh rule untuk meng-anotasi protein. Bagian kiri merupakan <kondisi> yang harus dipenuhi, misal apakah TAXON=Bacteria atau TAXON=Eukaryota. Jika semua kondisi tersebut terpenuhi, maka <konsekuensi> pada bagian kanan gambar akan dieksesusi,

misal protein tersebut akan dilabeli dengan terminologi yang sesuai (misal GO:0004742) dari Gene Ontology (Ashburner et al., 2000) termasuk fungsi (yang sudah diketahui) dari protein tersebut.



Gambar 26. Pelabelan data secara otomatis berbasis rule-based system (sumber: MacDougall et al., 2020)

Tabel berikut ini menunjukkan kumpulan sumber daya online yang dapat digunakan untuk melakukan pelabelan data, baik secara manual maupun otomatis.

Tabel 2. Metode Pelabelan Data Manual dan Otomatis

No	Keterangan	Link
1	VIA Annotator Software	https://www.robots.ox.ac.uk/~vgg/software/via/via.html
2	LabelMe Annotation Tool	http://labelme.csail.mit.edu/Release3.0/
3	TagLab	http://taglab.isti.cnr.it/
4	Snorkel	https://www.snorkel.org/
5	Natural Language Toolkit	https://www.nltk.org/
6	spaCy	https://spacy.io/

C. Rangkuman

Pelabelan data mempunyai peranan penting dalam kegiatan Analis Data Ilmiah yang bertujuan untuk memberikan informasi tambahan seperti konteks sehingga memungkinkan data untuk dimengerti dan diolah secara benar oleh manusia maupun mesin. Umumnya, pelabelan data dimaksudkan untuk digunakan dalam pembelajaran mesin terutama yang supervised. Dalam hal ini, data yang sudah dilabeli tersebut akan digunakan sebagai data latih, dimana masin akan mempelajari hubungan antara karakteristik data dengan label yang diberikan.

Secara spesifik, pelabelan data dapat digunakan antara lain untuk mengidentifikasi obyek-obyek tertentu dalam data berbentuk multimedia seperti gambar, audio, maupun video. Selain itu dapat juga digunakan untuk mengidentifikasi entitas serta hubungan antar entitas termasuk struktur yang ada. Ini jamak ditemukan dalam pengolahan bahasa alami seperti analisa sentimen dimana makna kata atau kalimat tidak saja ditentukan oleh arti dari kata atau kalimat itu sendiri, tetapi juga bagaimana kata dan kalimat tersebut berhungan dengan kata dan kalimat lainnya. Pelabelan data dapat juga digunakan untuk identifikasi semantik atau makna, dimana satu kata dapat merepresentasikan berbagai makna ataupun beberapa kata dapat mempunyai makna yang sama.

Untuk melakukan pelabelan data dapat menggunakan metode manual ataupun otomatis. Metode manual sangat tergantung pada kemampuan expert untuk memberikan tanda dan label, sedangkan metode otomatis mengandalkan kemampuan mesin untuk meniru kemampuan expert melakukan pelabelan. Dalam hal metode otomatis, kualitas pelabelan data sangat variatif tergantung pada teknik dan domain data.

D. Evaluasi

1. Identifikasi kasus kegiatan Analis Data Ilmiah dalam unit kerja atau institusi anda yang memerlukan dilakukannya pelabelan data yang berupa:
   a. Teks
   b. Gambar
   c. Audio
   d. Video

2. Dari setiap contoh kasus yang berhasil diidentifikasi pada nomor 1, jelaskan metode dan alat bantu pelabelan data yang paling sesuai.
3. Pilih salah satu dari contoh kasus yang diperoleh pada nomor 2 lalu lakukan pelabelan data (minimal 100 item data) menggunakan metode dan alat bantu yang diidentifikasi. Buat laporan yang setidaknya meliputi:
    a. Tujuan pelabelan data
    b. Bentuk data yang akan dilabeli
    c. Metode dan alat bantu pelabelan data yang digunakan
    d. Lama waktu rata-rata yang digunakan untuk melabeli 1 item data
    e. Total waktu yang dibutuhkan untuk melabeli keseluruhan item data

MATERI POKOK 3:

PRAPEMROSESAN DATA

Indikator Hasil Belajar:

Peserta mampu **melakukan pra-pemrosesan pada pengumpulan dan persiapan data dengan benar**, meliputi:

1. Menjelaskan prapemrosesan data dengan benar
2. Menjelaskan pentingnya kualitas data dengan benar
3. Melakukan validitas data dengan benar
4. Melakukan penanganan data yang dilang dengan benar
5. Melakukan data outlier/deteksi outlier dengan benar

A. Pengenalan Prapemrosesan Data

Data dalam bentuk asli (data dunia nyata) biasanya belum siap untuk digunakan dalam tugas analitik. Data sering kotor, tidak selaras, terlalu rumit, dan tidak akurat. Prapemrosesan Data (disebut juga *preprocessing data*) diperlukan untuk mengubah data mentah dunia nyata menjadi bentuk yang disempurnakan dengan baik untuk algoritma analitik (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Banyak profesional analitik akan bersaksi bahwa waktu yang dihabiskan untuk pra-pemrosesan data (yang mungkin merupakan fase yang paling tidak menyenangkan dalam keseluruhan proses) secara signifikan lebih lama daripada waktu yang dihabiskan untuk tugas analitik lainnya (kegembiraan dalam membangun dan penilaian model analitik).

Tujuan prapemrosesan data (lebih umum disebut *data preprocessing*) adalah untuk mengambil data yang diidentifikasi pada langkah sebelumnya dan mempersiapkannya untuk dianalisis dengan metode data mining. Dibandingkan dengan langkah-langkah lain dalam data mining, prapemrosesan data memakan waktu dan tenaga paling banyak; sebagian besar percaya bahwa langkah ini menyumbang sekitar 80% dari total waktu yang dihabiskan untuk proyek data mining. Alasan untuk upaya yang sangat besar yang dihabiskan pada langkah ini adalah fakta bahwa data yang didapatkan umumnya tidak lengkap (*missing*

value), seperti kurang nilai atribut, kurang atribut tertentu yang menarik, atau hanya berisi data agregat), data *noise* (mengandung kesalahan atau outlier), dan tidak konsisten (mengandung perbedaan dalam kode atau nama).

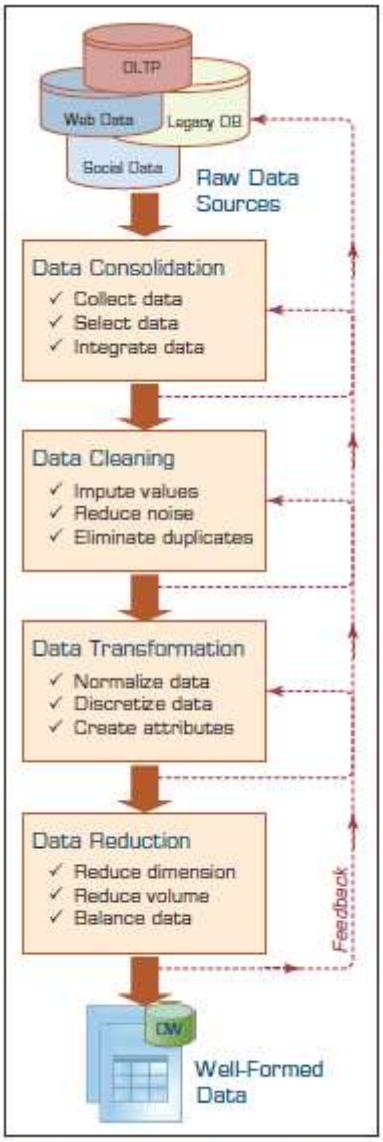


FIGURE 2.3 Data Preprocessing Steps.

Gambar 27. Langkah-langkah utama dalam upaya preprocessing data

Pada fase pertama Prapemrosesan Data, data yang relevan dikumpulkan dari sumber yang teridentifikasi, catatan dan variabel yang diperlukan dipilih (berdasarkan pemahaman yang mendalam tentang data, informasi yang tidak perlu disaring), dan catatan datang dari berbagai sumber data diintegrasikan/ digabungkan.

Pada fase kedua Prapemrosesan Data, data dibersihkan (*data cleaning*). Data dalam bentuk aslinya/mentah/dunia nyata biasanya kotor (Hernández & Stolfo, 1998; Kim et al., 2003). Pada langkah ini, nilai-nilai dalam kumpulan data diidentifikasi dan ditangani. Dalam beberapa kasus, nilai yang hilang (*missing value*) merupakan anomali dalam kumpulan data, dalam hal ini nilai tersebut perlu diperhitungkan (diisi dengan nilai yang paling mungkin) atau diabaikan; dalam kasus lain, nilai yang hilang adalah bagian alami dari kumpulan data (misalnya, bidang pendapatan rumah tangga sering tidak dijawab oleh orang-orang yang berada di tingkat pendapatan teratas). Pada langkah ini, analis juga harus mengidentifikasi nilai *noise* dalam data (yaitu outlier) dan memuluskannya. Selain itu, inkonsistensi (nilai yang tidak biasa dalam suatu variabel) dalam data harus ditangani dengan menggunakan pengetahuan domain dan/atau pendapat ahli.

Pada fase ketiga preprocessing data, data diubah untuk pemrosesan yang lebih baik. Misalnya, dalam banyak kasus data dinormalisasi antara minimum dan maksimum tertentu untuk semua variabel untuk mengurangi potensi bias dari satu variabel (memiliki nilai numerik yang besar, seperti pendapatan rumah tangga) yang mendominasi variabel lain (seperti jumlah tanggungan atau tahun). dalam layanan, yang berpotensi menjadi lebih penting) memiliki nilai yang lebih kecil. Transformasi lain yang terjadi adalah diskritisasi dan/atau agregasi. Dalam beberapa kasus, variabel numerik dikonversi menjadi nilai kategorikal (misalnya, rendah, sedang, tinggi); dalam kasus lain, rentang nilai unik variabel nominal dikurangi menjadi set yang lebih kecil menggunakan hierarki konsep (misalnya, sebagai lawan dari menggunakan masing-masing negara bagian dengan 50 nilai berbeda, seseorang dapat memilih untuk menggunakan beberapa wilayah untuk variabel yang menunjukkan lokasi) untuk dimiliki kumpulan data yang lebih dapat menerima pemrosesan komputer. Namun, dalam kasus lain seseorang mungkin memilih untuk membuat variabel baru berdasarkan yang sudah ada untuk memperbesar informasi yang ditemukan dalam kumpulan variabel dalam kumpulan data. Misalnya, dalam kumpulan data transplantasi organ, seseorang dapat memilih untuk menggunakan variabel tunggal yang menunjukkan kecocokan golongan darah (1: cocok, 0: tidak cocok) dibandingkan dengan nilai multinominal terpisah untuk golongan darah donor dan donor. penerima.

Penyederhanaan tersebut dapat meningkatkan konten informasi sekaligus mengurangi kompleksitas hubungan dalam data.

Tahap akhir dari preprocessing data adalah reduksi data. Meskipun ilmuwan data (*data scientists*) ingin memiliki kumpulan data yang besar, terlalu banyak data juga dapat menjadi masalah. Dalam pengertian yang paling sederhana, seseorang dapat memvisualisasikan data yang biasa digunakan dalam proyek analitik prediktif sebagai file datar yang terdiri dari dua dimensi: variabel (jumlah kolom) dan kasus/rekaman (jumlah baris). Dalam beberapa kasus (misalnya, pemrosesan gambar dan proyek genom dengan data microarray yang kompleks), jumlah variabel bisa lebih besar, dan analis harus mengurangi jumlahnya menjadi ukuran yang dapat dikelola. Karena variabel diperlakukan sebagai dimensi berbeda yang menggambarkan fenomena dari perspektif berbeda, dalam analitik prediktif dan penambangan data, proses ini biasa disebut reduksi dimensi (atau pemilihan variabel). Meskipun tidak ada satu pun cara terbaik untuk menyelesaikan tugas ini, seseorang dapat menggunakan temuan dari literatur yang diterbitkan sebelumnya; berkonsultasi dengan pakar domain; menjalankan uji statistik yang sesuai (misalnya, analisis komponen utama atau analisis komponen independen); dan, yang lebih disukai, gunakan kombinasi teknik ini untuk berhasil mengurangi dimensi dalam data menjadi subset yang lebih mudah dikelola dan paling relevan.

B. Pentingnya Kualitas Data

Data adalah bahan utama untuk inisiatif *Business Intelligence*, ilmu data, dan analitik bisnis apa pun. Faktanya, data dapat dilihat sebagai bahan mentah untuk apa yang dihasilkan oleh teknologi keputusan populer ini—informasi, wawasan, dan pengetahuan. Tanpa data, tidak satu pun dari teknologi ini dapat ada dan dipopulerkan—walaupun, secara tradisional kami telah membangun model analitik menggunakan pengetahuan dan pengalaman ahli ditambah dengan sangat sedikit atau tanpa data sama sekali; namun, itu dulu, dan sekarang data adalah intinya. Setelah dianggap sebagai tantangan besar untuk mengumpulkan, menyimpan, dan mengelola, data saat ini secara luas dianggap sebagai aset organisasi yang paling berharga, dengan potensi untuk menciptakan wawasan yang tak ternilai untuk lebih memahami pelanggan, pesaing, dan proses bisnis.

Data bisa kecil atau bisa sangat besar. Data dapat berupa terstruktur (diatur dengan baik untuk diproses oleh komputer), atau dapat tidak terstruktur (misalnya, teks yang dibuat untuk manusia dan karenanya tidak mudah dimengerti/ dikonsumsi oleh komputer). Data bisa datang dalam *batch* yang lebih kecil terus menerus atau bisa dituangkan sekaligus sebagai *batch* besar. Ini adalah beberapa karakteristik yang menentukan sifat bawaan dari data saat ini, yang sering kita sebut Big Data. Meskipun karakteristik data ini membuatnya lebih menantang untuk diproses dan dikonsumsi, namun juga membuatnya lebih berharga karena memperkaya data di luar batas konvensionalnya, memungkinkan penemuan pengetahuan baru dan baru. Cara tradisional untuk mengumpulkan data secara manual (baik melalui survei atau melalui transaksi bisnis yang dimasuki manusia) sebagian besar meninggalkan tempat mereka ke mekanisme pengumpulan data modern yang menggunakan Internet dan/atau jaringan komputerisasi berbasis sensor/RFID. Sistem pengumpulan data otomatis ini tidak hanya memungkinkan untuk mengumpulkan lebih banyak volume data tetapi juga meningkatkan kualitas dan integritas data.

Kualitas data menjadi perhatian yang berkelanjutan di mana pun data dikumpulkan, diproses, dan disimpan. Dalam kumpulan data yang ditampilkan pada tabel 3 berikut.

Tabel 3. Kumpulan data

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

Bagaimana kita mengetahui apakah data skor kredit dan suku bunga akurat? Bagaimana jika skor kredit memiliki nilai tercatat 900 (di luar batas teoretis) atau jika ada kesalahan entri data? Kesalahan dalam data ini akan memengaruhi keterwakilan model. Organisasi menggunakan teknik pembersihan dan transformasi data untuk meningkatkan dan mengelola kualitas data dan

menyimpannya di repositori seluruh perusahaan yang disebut *Data Warehouse* (Gudang Data). Data yang bersumber dari gudang data yang terpelihara dengan baik memiliki kualitas yang lebih tinggi, karena terdapat kontrol yang tepat untuk memastikan tingkat akurasi data untuk data baru dan yang sudah ada. Praktik pembersihan data meliputi penghapusan catatan duplikat, mengkarantina catatan outlier yang melebihi batas, standarisasi nilai atribut, penggantian nilai yang hilang, dll. Terlepas dari itu, sangat penting untuk memeriksa data menggunakan teknik eksplorasi data selain menggunakan pengetahuan sebelumnya tentang data dan bisnis sebelum membangun model untuk memastikan tingkat kualitas data tertentu.

C. Validasi Data

Kualitas input data mungkin terbukti tidak memuaskan karena ketidaklengkapan (*Incompleteness*), gangguan (*noise*), dan ketidakkonsistenan (*inconsistency*).

1. Ketidaklengkapan (*Incompleteness*)

Beberapa rekaman mungkin berisi nilai yang hilang terkait dengan satu atau beberapa atribut, dan mungkin ada berbagai alasan untuk hal ini. Mungkin beberapa data tidak dicatat pada sumbernya secara sistematis, atau tidak tersedia saat transaksi yang terkait dengan catatan terjadi. Dalam kasus lain, data mungkin hilang karena alat perekam tidak berfungsi. Mungkin juga beberapa data sengaja dihapus selama tahap sebelumnya dari proses pengumpulan karena dianggap tidak benar. Ketidaklengkapan juga dapat berasal dari kegagalan untuk mentransfer data dari database operasional ke data mart yang digunakan untuk analisis intelijen bisnis tertentu.

2. Gangguan (*noise*)

Data mungkin mengandung nilai yang salah atau anomali, yang biasanya disebut sebagai *outlier*. Kemungkinan penyebab *noise* lainnya harus dicari di perangkat yang tidak berfungsi untuk pengukuran, perekaman, dan transmisi data. Adanya data yang dinyatakan dalam unit pengukuran yang heterogen, yang karenanya memerlukan konversi, pada gilirannya dapat menyebabkan anomali dan ketidakakuratan.

3. **Ketidakkonsistenan (*inconsistency*)**

   Terkadang data mengandung ketidaksesuaian karena perubahan dalam sistem pengkodean yang digunakan untuk representasi data sehingga mungkin tampak tidak konsisten. Misalnya, pengkodean produk yang diproduksi oleh perusahaan dapat mengikuti pada revisi yang berlaku pada tanggal tertentu, tanpa data yang direkam pada periode sebelumnya mengikuti pada transformasi yang diperlukan untuk menyesuaikannya dengan skema pengkodean yang direvisi.

Tujuan dari teknik validasi data adalah untuk mengidentifikasi dan menerapkan tindakan korektif jika ada data yang tidak lengkap dan tidak konsisten atau data yang terkena *noise*.

D. Penanganan Data yang Hilang (*Missing Value*)

Salah satu masalah kualitas data yang paling umum adalah bahwa beberapa rekaman memiliki nilai atribut yang hilang. Misalnya, skor kredit mungkin hilang di salah satu catatan. Ada beberapa metode mitigasi yang berbeda untuk mengatasi masalah ini, namun masing-masing metode memiliki pro dan kontra. Langkah pertama dalam mengelola nilai yang hilang adalah memahami alasan di balik hilangnya nilai tersebut. Melacak silsilah data dari sumber data dapat mengarah pada identifikasi masalah sistemik dalam pengambilan data, kesalahan dalam transformasi data, atau mungkin ada fenomena yang belum dipahami pengguna. Mengetahui sumber nilai yang hilang seringkali akan memandu metodologi mitigasi apa yang akan digunakan. Kita dapat mengganti nilai yang hilang dengan rentang data buatan sehingga kita dapat mengelola masalah dengan dampak kecil pada langkah selanjutnya dalam penambangan data. Nilai skor kredit yang hilang dapat diganti dengan skor kredit yang berasal dari kumpulan data (nilai rata-rata atau minimum atau maksimum, bergantung pada karakteristik atribut). Metode ini berguna jika nilai yang hilang terjadi secara acak dan frekuensi kemunculannya cukup jarang. Jika tidak, distribusi atribut yang memiliki data yang hilang akan terdistorsi. Alternatifnya, untuk membangun model representatif, kita dapat mengabaikan semua catatan data dengan nilai yang hilang atau catatan dengan kualitas data yang buruk. Metode ini mengurangi ukuran kumpulan data. Beberapa algoritme penambangan data

bagus dalam menangani catatan dengan nilai yang hilang, sementara yang lain mengharapkan langkah persiapan data untuk menanganinya sebelum model dibangun dan diterapkan. Sebagai contoh, algoritma k-nearest neighbor (k-NN) untuk tugas klasifikasi seringkali kuat dengan nilai yang hilang. Model jaringan saraf untuk tugas klasifikasi tidak bekerja dengan baik dengan atribut yang hilang dan dengan demikian langkah persiapan data sangat penting untuk mengembangkan model jaringan saraf.

Untuk mengoreksi sebagian data yang tidak lengkap, seseorang dapat menggunakan beberapa teknik berikut:

1. **Eliminasi**

   Dimungkinkan untuk membuang semua record yang nilai dari satu atau lebih atributnya hilang. Dalam kasus analisis penambangan data yang diawasi, sangat penting untuk menghilangkan catatan jika nilai atribut target hilang. Kebijakan yang didasarkan pada penghapusan catatan secara sistematis mungkin tidak efektif ketika distribusi nilai yang hilang bervariasi dengan cara yang tidak teratur di seluruh atribut yang berbeda, karena seseorang dapat menanggung risiko kehilangan informasi yang substansial.

2. **Inspeksi**

   Sebagai alternatif, seseorang dapat memilih inspeksi dari setiap nilai yang hilang, yang dilakukan oleh para ahli di domain aplikasi, untuk mendapatkan rekomendasi tentang kemungkinan nilai pengganti. Jelas, pendekatan ini menderita kesewenang-wenangan dan subjektivitas tingkat tinggi, dan agak memberatkan dan memakan waktu untuk kumpulan data besar. Di sisi lain, pengalaman menunjukkan bahwa itu adalah salah satu tindakan korektif paling akurat jika dilakukan dengan terampil.

3. **Identifikasi**

   Sebagai kemungkinan ketiga, nilai konvensional dapat digunakan untuk menyandikan dan mengidentifikasi nilai yang hilang, sehingga tidak perlu menghapus seluruh rekaman dari kumpulan data yang diberikan. Misalnya, untuk atribut kontinu yang hanya mengasumsikan nilai positif, dimungkinkan untuk menetapkan nilai {−1} ke semua data yang hilang. Dengan cara yang sama, untuk atribut kategoris seseorang dapat mengganti nilai yang hilang dengan nilai baru yang berbeda dari semua yang diasumsikan oleh atribut.

4. **Pengganti**.

   Ada beberapa kriteria untuk penggantian otomatis data yang hilang, meskipun sebagian besar dari mereka tampak sewenang-wenang. Misalnya, nilai atribut yang hilang dapat diganti dengan rata-rata atribut yang dihitung untuk pengamatan yang tersisa. Teknik ini hanya dapat diterapkan pada atribut numerik, tetapi jelas tidak akan efektif dalam kasus distribusi nilai yang asimetris. Dalam analisis terbimbing juga dimungkinkan untuk mengganti nilai yang hilang dengan menghitung rata-rata atribut hanya untuk record yang memiliki kelas target yang sama. Terakhir, nilai kemungkinan maksimum, yang diestimasi menggunakan model regresi atau metode Bayesian, dapat digunakan sebagai pengganti nilai yang hilang. Namun, prosedur perkiraan dapat menjadi agak rumit dan memakan waktu untuk kumpulan data besar dengan persentase data yang hilang yang tinggi.

E. Standarisasi/Normalisasi Data

Dalam sebagian besar analisis data mining, adalah tepat untuk menerapkan beberapa transformasi pada kumpulan data untuk meningkatkan akurasi model pembelajaran yang dikembangkan selanjutnya. Memang, teknik koreksi outlier adalah contoh transformasi data asli yang memfasilitasi fase pembelajaran selanjutnya. Metode principal component juga dapat dianggap sebagai proses transformasi data.

Sebagian besar model pembelajaran mendapat manfaat dari standarisasi pencegahan data, juga disebut normalisasi. Teknik standardisasi yang paling populer meliputi metode penskalaan desimal, metode min-max, dan metode indeks-z.

1. **Skala desimal**

   Penskalaan desimal didasarkan pada transformasi:

$$x'_{ij} = \frac{x_{ij}}{10^h},$$

   di mana h adalah parameter tertentu yang menentukan intensitas penskalaan. Dalam praktiknya, penskalaan desimal sama dengan menggeser titik desimal dengan posisi h ke arah kiri. Secara umum, h ditetapkan pada nilai yang memberikan nilai transformasi dalam rentang [−1, 1].

2. Min-maks

Standardisasi min-max dicapai melalui transformasi:

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}(x'_{\max,j} - x'_{\min,j}) + x'_{\min,j},$$

dimana:

$$x_{\min,j} = \min_i x_{ij}, \quad x_{\max,j} = \max_i x_{ij},$$

adalah nilai minimum dan maksimum dari atribut $\mathbf{a}_j$ sebelum transformasi, sedangkan $x'_{\min,j}$ dan $x'_{\max,j}$ adalah nilai minimum dan maksimum yang ingin kita peroleh setelah transformasi. Secara umum, nilai ekstrim dari jangkauan didefinisikan sehingga

$x'_{\min,j} = -1$ and $x'_{\max,j} = 1$ or $x'_{\min,j} = 0$ and $x'_{\max,j} = 1$

3. z -indeks

Standardisasi berbasis z-index menggunakan transformasi:

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

di mana μ¯j dan σ¯j masing-masing adalah rata-rata sampel dan deviasi standar sampel dari atribut aj . Jika distribusi nilai atribut aj kira-kira normal, transformasi berbasis indeks-z menghasilkan nilai yang hampir pasti berada dalam rentang (−3, 3).

F. Data Outlier/Deteksi Outlier

Outlier menurut definisi adalah anomali dalam kumpulan data. Penyimpangan dapat terjadi secara sah (pendapatan dalam miliaran) atau keliru (tinggi manusia 1,73 sentimeter). Terlepas dari itu, keberadaan outlier perlu dipahami dan membutuhkan perlakuan khusus. Tujuan pembuatan model representatif adalah untuk menggeneralisasikan pola atau hubungan dalam data dan adanya outlier yang mencondongkan model. Teknik pendeteksian outlier akan dibahas secara detail pada Bab 11 Deteksi Anomali pada deteksi anomali. Mendeteksi outlier mungkin menjadi tujuan utama dari beberapa aplikasi penambangan data, seperti deteksi penipuan dan deteksi intrusi.

G. Rangkuman

Inti dari prapemrosesan data data dirangkum dalam Tabel berikut, yang memetakan fase utama (bersama dengan deskripsi masalahnya) ke daftar tugas dan algoritme yang representatif.

Proses Utama	Sub Proses	Metode Popular yang Digunakan
Data consolidation	Mengakses dan mengumpulkan data	SQL queries, software agents, Web services
	Pilih dan filter data	Domain expertise, SQL queries, statistical tests
	Mengintegrasikan dan menyatukan data	SQL queries, domain expertise, ontology-driven data mapping
Data cleaning	Menangani nilai yang hilang (missing value) dalam data	Isikan nilai yang hilang imputations) dengan nilai yang paling sesuai (rata-rata, median, min/maks, modus, dll.); recode nilai yang hilang dengan konstanta seperti "ML"; hapus catatan nilai yang hilang; tidak melakukan apapun.
	Mengidentifikasi dan mengurangi noise pada data	Identifikasi outlier dalam data dengan teknik statistik sederhana (seperti rata-rata dan standar deviasi) atau dengan analisis kluster; setelah diidentifikasi, hapus outlier atau ratakan dengan menggunakan *binning*, regresi, atau rata-rata sederhana.
	Temukan dan hilangkan data yang salah	Identifikasi nilai yang salah dalam data (selain outlier), seperti nilai ganjil, label kelas yang tidak konsisten, distribusi ganjil; setelah

		teridentifikasi, gunakan keahlian domain untuk memperbaiki nilai atau menghapus catatan yang menyimpan nilai yang salah.
Data transformation	Normalisasi Data	Kurangi rentang nilai di setiap variabel bernilai numerik ke rentang standar (mis., 0 hingga 1 atau –1 hingga +1) dengan menggunakan berbagai teknik normalisasi atau penskalaan
	Diskritisasi atau agregat data	Jika perlu, ubah variabel numerik menjadi representasi diskrit menggunakan teknik binning berbasis rentang atau frekuensi; untuk variabel kategori, kurangi jumlah nilai dengan menerapkan hierarki konsep yang tepat
	Membangun atribut baru	Turunkan variabel baru dan lebih informatif dari yang sudah ada menggunakan berbagai fungsi matematika (sederhana seperti penjumlahan dan perkalian atau serumit kombinasi hibrid dari transformasi log).
Data reduction	Kurangi jumlah atribut	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Kurangi jumlah record	Random sampling, stratified sampling,

		expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample kelas yang kurang terwakili atau undersample kelas yang lebih terwakili

H. Evaluasi

1. Mengapa data asli/mentah tidak siap digunakan oleh tugas analitik?
2. Apa langkah-langkah preprocessing data utama?
3. Apa yang dimaksud dengan membersihkandata? Kegiatan apa yang dilakukan pada fase ini?
4. Mengapa kita membutuhkan transformasi data? Apa tugas transformasi data yang umum digunakan?
5. Mengapa deteksi outlier itu penting?

MATERI POKOK 4:

PEREKAYASAAN FITUR

Indikator Hasil Belajar:
Peserta mampu **mengidentifikasi fitur data untuk pengolahan data dengan benar**, meliputi:

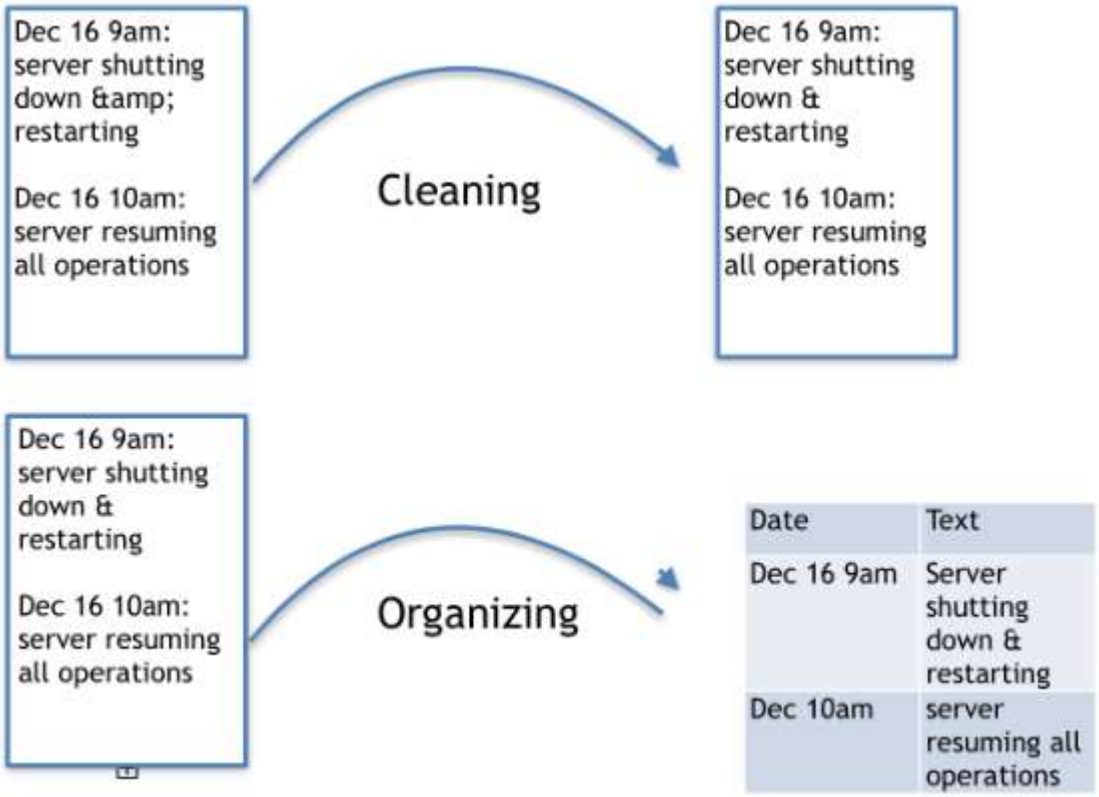
1. Mengidentifikasi definisi dan ruang lingkup fitur data untuk pengolahan data dengan benar
2. Mengidentifikasi seleksi fitur data untuk pengolahan data dengan benar
3. Mengidentifikasi transformasi fitur data untuk pengolahan data dengan benar

A. Defenisi dan Ruang Lingkup

Terminologi perekayaan fitur (*feature engineering*) umumnya berkaitan dengan teknik pembelajaran mesin (*machine learning*). Dalam kasus tersebut, perekayasaan fitur dimaknai sebagai proses untuk memilih, mengekstrak, menggabungkan, mentransformasi fitur-fitur dari data (utamanya data mentah) sehingga dapat digunakan oleh algoritma pembelajaran mesin. Berbagai macam fitur dapat diperoleh seperti karakteristik, properti, attribute dari data. Proses ini masuk dalam tahapan prapemrosesan data yaitu mempersiapkan fitur yang tepat sehingga dapat dihasilkan hasil pembelajaran mesin yang lebih optimal. Namun demikian, perbedaan utama dengan proses-proses prapemrosesan data yang dijelaskan pada materi sebelumnya, proses perekayasaan fitur umumnya melakukan transformasi data yang komprehensif misalnya merubah seluruh format data kedalam bentuk yang lebih bermanfaat untuk proses pembelajaran.

Gambar 28 menunjukkan contoh perbedaan antara salah satu metode prapemrosesan data (*cleaning*) dan proses perekayasaan fitur (*organizing*). Pada metode prapemrosesan data tersebut, data yang berupa akses log server akan ditransformasi sehingga sintaks dalam bahasa mesin berubah kedalam bentuk yang lebih dimengerti (misalnya “&” dibersihkan menjadi “&”). Pada proses perekayasaan fitur, selain melakukan proses pembersihan tersebut, fitur penting dari data akan dipilih dan diekstrak (misal teks “Dec 16 9am” akan dikenali sebagai attribut data berupa Tanggal). Selanjutnya, data dapat

ditransformasi dalam bentuk tabular yang terdiri dari 2 attribut yaitu “Date” dan “Text”. Pada akhirnya, hasil transformasi data tersebut akan menunjukkan format data yang lebih informatif dan bermanfaat dalam melakukan pembelajaran mesin.



Gambar 28. Perbedaan antara proses prapemrosesan data (cleaning) dan perekayasaan fitur (organizing) (sumber: Ozdemir & Susarla, 2018)

Berdasarkan illustrasi diatas, maka perekayasaan fitur dalam Analis Data Ilmiah dapat didefinisikan sebagai proses untuk mentransformasi data menjadi fitur-fitur yang merepresentasikan permasalahan yang ingin diselesaikan secara lebih baik sehingga menghasilkan pelaksanaan pembelajaran mesin yang lebih baik. Dengan demikian, proses perekayasaan fitur mempunyai ciri sebagai berikut:

1. Merupakan proses transformasi data. Dalam hal ini, data tidak harus tidak berupa data mentah. Perekayasaan fitur dapat diterapkan pada berbagai level data termasuk data yang sudah diproses sekalipun. Namun pada umumnya, proses ini diterapkan pada data mentah yang sudah diberi perlakuan proses prapemrosesan sebagaimana dijelaskan pada materi sebelumnya.

2. Fitur. Merupakan attribut data yang mempunyai makna atau signifikan pada proses pembelajaran mesin. Penting diingat bahwa makna atau signifikansi dari attribut tersebut ditentukan oleh tujuan pembelajaran mesin. Sebagai contoh, jika kita mempunyai data demografi penduduk dengan attribut nama, jenis kelamin, umur, tingkat pendidikan, pekerjaan dst. Jika tujuan pembelajaran mesin adalah memetakan potensi peserta sosialisasi tentang social media, maka jika diasumsikan anak muda mempunyai tingkat ketertarikan yang lebih tinggi pada topik tersebut, maka attribut umur akan mempunyai tingkat signifikansi yang lebih tinggi.
3. Fitur mempunyai kemampuan yang lebih baik dalam merepresentasikan permasalahan yang ingin diselesaikan. Transformasi data yang dilakukan dimaksudkan untuk memperoleh representasi permasalahan dengan lebih baik.
4. Proses perekayasaan fitur memperbaiki hasil pembelajaran mesin yang dilakukan. Sebagai bagian dari tahapan persiapan data, proses ini tidak terpisahkan dari tahapan analisis data. Keberhasilan proses ini tentu saja akan ditentukan oleh proses analisis data nantinya, yang dalam hal pembelajaran mesin dapat diukur melalui tingkat akurasi yang lebih tinggi atau tingkat error yang lebih rendah atau lama waktu analisis yang lebih singkat. Dengan demikian, perlu dipahami bahwa proses perekayasaan fitur semestinya tidak berhenti pada dihasilkannya fitur data tetapi terpakai dalam proses pembelajaran mesin di tahapan analisis data.

B. Memahami Fitur

Sebagaimana diuraikan pada bagian sebelumnya, fitur adalah attribut data yang mempunyai pengaruh signifikan dalam proses pembelajaran mesin. Untuk dapat memahami attribut data tersebut, perlu diketahui karakteristik data sebagai berikut:

1. Data dapat dibedakan berdasarkan strukturnya, yaitu data yang terstruktur dan data tidak terstruktur. Pada data yang terstruktur, data dapat dipecah antara karakteristik dan observasi. Dalam format tabular, karakteristik berupa kolom dan obervasi berupa baris. Sedangkan data yang tidak terstruktur tidak mempunyai standar tertentu dan biasanya menggunakan satu karakterikstik

(kolom) saja.

2. Data dapat dibedakan berdasarkan tipenya, yaitu data quantitative dan data qualitative. Data quantitative mempunyai bentuk numerikal sedangkan data qualitative berbentuk kategorikal.
3. Data dapat dibedakan berdasarkan tingkatannya dimana tiap tingkatan mempunyai batasan operasi yang boleh dilakukan:
   a. Nominal, merupakan tingkatan data yang paling sederhana dimana data ditandai dengan nama semata. Misalnya, nama orang, golongan darah, jenis binatang, dll. Tipe data berbentuk kualitatif. Pada data jenis ini tidak dapat dilakukan operasi matematika kuantitatif seperti penjumlahan atau pembagian.
   b. Ordinal, data dalam jenis ini mempunyai urutan secara alami sehingga dapat diasumsikan bahwa ada item data yang mempunyai nilai lebih baik dari item lainnya. Tipe data berbentuk kualitatif. Contoh data yaitu ranking kelas, rating acara televisi, atau nilai ujian. Pada data jenis ini dapat dilakukan operasi matematika seperti menghitung nilai tengah (median) atau persentil.
   c. Interval, data dalam jenis ini mempunyai urutan sebagaimana data ordinal tetapi lebih dari itu dapat diketahui juga perbedaan yang berarti diantara nilai-nilai. Tipe data berbentuk kuantitatif. Contoh data yaitu temperatur, dimana jika diketahui temperatur di Jakarta adalah 34 derajat Celcius dan di Bogor 30 derajat Celcius, maka kita dapat menghitung perbedaan temperatur diantara kedua tempat (34-30 = 4 derajat Celcius). Operasi matematika yang diperbolehkan pada data jenis ini antara lain penjumlahan, pengurangan, menghitung nilai rata-rata ataupun standar deviasi.
   d. Ratio, sama dengan data jenis Interval, tipe data jenis ini berbentuk kuantitatif dengan tambahan adanya nilai nol mutlak. Akibatnya, selain dapat dilakukan penjumlahan dan pengurangan, data jenis ini dapat diperkalikan dan dibagi. Contoh data adalah nilai mata uang, dimana uang senilai Rp. 1000 adalah dua kali lebih besar dari Rp. 500 sebab 1000/500 = 2.

Pemahaman atas attribut data (fitur) tersebut menjadi langkah awal dalam melakukan rekayasa fitur terutama karena setiap jenis fitur mempunyai kemampuan yang berbeda dalam menerima operasi matematika. Sebagai contoh, jika diperoleh dataset yang baru, seorang analis data perlu melakukan identifikasi tipe data untuk setiap fiturnya, kemudian tingkatannya sehingga dapat ditentukan jenis rekayasa yang bisa dilakukan, misalnya jenis transformasi melalui operasi matematis yang bisa dilakukan.

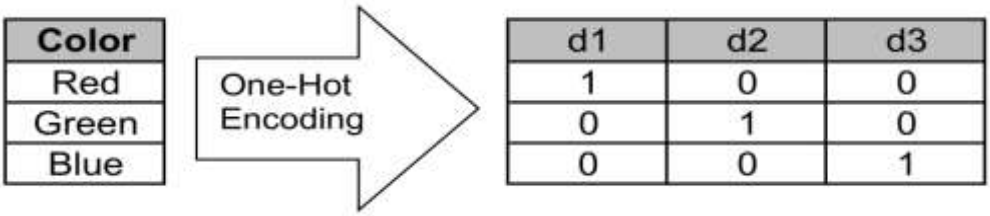
1. Pengembangan Fitur: Data Kategorikal

Hal umum yang ditemukan dalam proses perekayasaan fitur adalah data yang mempunyai tipe kualitatif dimana data berbentuk kategorikal. Sebagaimana pada penjelasan dibagian sebelumnya, data berbentuk kategorikal mempunyai keterbatasan dalam hal operasi yang bisa dilakukan. Bahkan tipe data kualitatif dengan level nominal tidak bisa dilakukan operasi matematis kuantitatif. Sesuai dengan tujuan proses perekayasaan fitur yang disampaikan diawal materi ini, attribut data dapat ditransformasi sehingga mempunyai kemampuan yang baik dalam proses pembelajaran mesin nantinya.

Salah satu metode yang umum digunakan untuk mentransformasi data kategorikal ke numerikal adalah melalui proses yang disebut enkoding (*encoding*). Terdapat berbagai metode encoding, antara lain dijelaskan berikut ini:

a. One-Hot Encoding

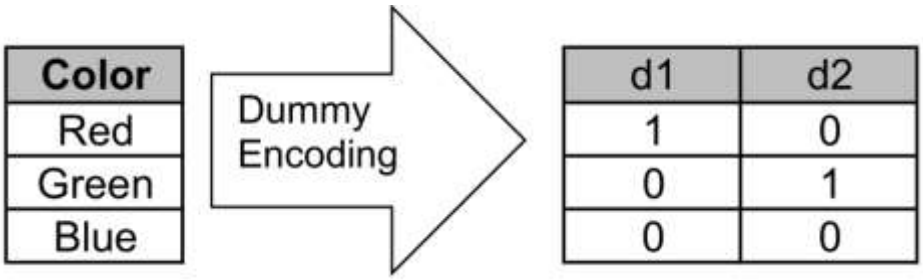
One-hot encoding adalah metode untuk mentransformasi data kategorikal menjadi numerikal dengan cara memberikan nilai untuk setiap kategori unik dalam suatu fitur. Contoh proses ini digambarkan pada Gambar 2, dimana data asli mempunyai 1 attribut yaitu “Color” yang mempunyai nilai dengan 3 kategori unik yaitu “Red”, “Green” dan “Blue”. Selanjutnya, dibuat attribut baru sebanyak jumlah kategori unik tersebut (dalam hal ini 3). Untuk setiap kategori, nilai attribut yang bersesuaian akan diberi nilai numerik “1” sedangkan fitur yang lain akan bernilai “0”. Dari contoh ini, kategori “Red” akan menjadi “100”, “Green” akan menjadi “010”, serta “Blue” menjadi “001”.



Gambar 29. Illustrasi proses one-hot encoding

b. Dummy Encoding

Proses dummy encoding hampir sama dengan one-hot encoding terkecuali pada jumlah fitur yang dihasilkan. Jika jumlah nilai kategori unik dari data adalah N, maka one-hot encoding akan menghasilkan N fitur, sedangkan dummy encoding akan menghasilkan N-1 fitur. Dummy encoding akan membuat fitur “dummy” dengan nilai “0” untuk merepresentasikan nilai kategori lainnya. Gambar 3 mengilustrasikan proses dummy encoding dengan nilai kategori unik sebanyak 3. Dihasilkan 2 fitur, dimana nilai untuk kategori terakhir adalah “00”. Sebagai hasil, ketegori “Red” akan menjadi “10”, “Green” menjadi “01” sedangkan fitur dummy “Blue” akan menjadi “00” yang berarti selain “Red” atau “Green”.



Gambar 30. Illustrasi proses dummy-encoding

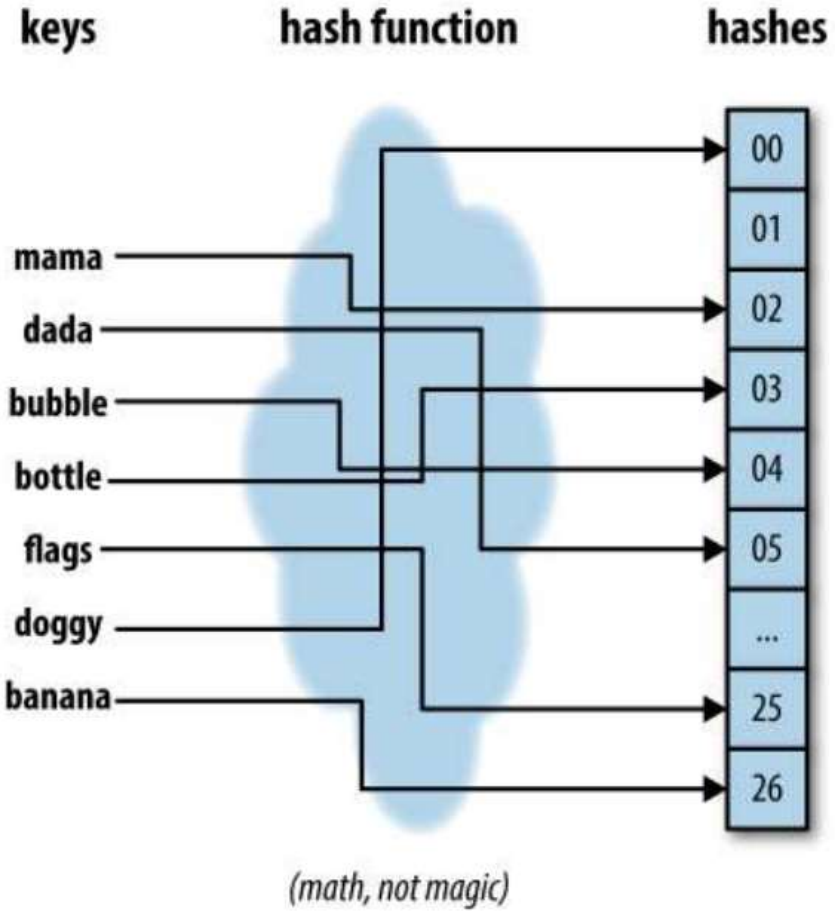
Terdapat juga metode serupa dengan dummy encoding yang disebut effect encoding dengan perbedaan kecil. Pada dummy encoding digunakan 1 dan 0 untuk merepresentasikan data, sedangkan effect encoding menggunakan 1, 0 dan -1.

c. Hash Encoding

Metode encoding yang diperkenalkan sebelumnya dapat digunakan

jika jumlah nilai kategorikal yang unik dari data relatif kecil. Jika jumlahnya besar, maka metode enkoding tersebut menjadi tidak efisien, misalnya akan menggunakan memori komputasi yang cukup besar pula. Salah satu metode encoding yang bisa digunakan untuk mengatasi hal tersebut adalah hash encoding.

Gambar 31 mengillustrasikan proses hash encoding, dimana fungsi hash adalah fungsi deterministik yang memetakan sejumlah input (yang bisa saja berukuran besar) ke rentang nilai tertentu. Proses encoding ini juga dapat dimaknai merupakan pemetaan input ke kelompok output (*bins*), dimana jumlah yang dipetakan ke setiap bin adalah relatif sama.



Gambar 31. Illustrasi proses hash encoding (sumber: Zheng & Casari, 2018)

2. Pengembangan Fitur: Data Tekstual

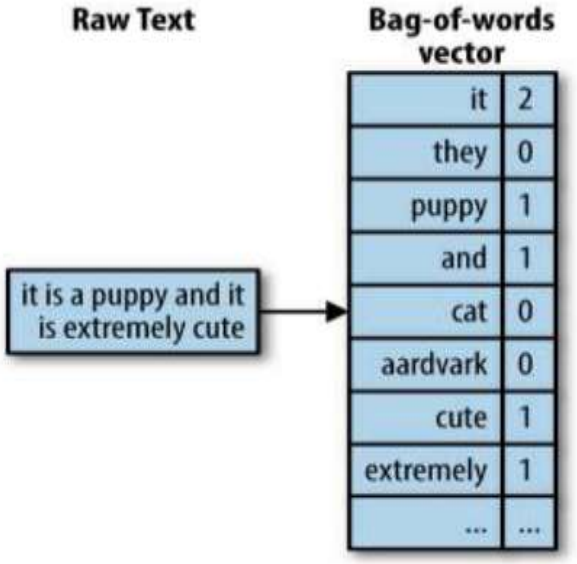
Tantangan terkait perekayasaan fitur lainnya adalah data tidak terstruktur yang berbentuk teks. Data jenis ini banyak ditemukan di Internet, misalnya

tulisan berita online, posting sosial media, review jasa ataupun produk. Meskipun format datanya tidak terstruktur, tetapi data jenis ini kaya akan informasi yang dapat dianalisa menggunakan pembelajaran mesin. Sebagai contoh, posting sosial media dapat digunakan untuk mengukur sentimen masyarakat terhadap layanan jasa atau produk tertentu, review online bisa digunakan untuk mengetahui hotel atau restoran yang terbaik yang bisa dikunjungi.

Untuk proses pengolahan data tekstual ini, dikenal metode Natural Language Processing (NLP) dimana komputer diprogram secara spesifik untuk mengolah data manusia. Ruang lingkup ilmu NLP sangat luas, sehingga pada pembahasan pengembangan fitur ini dibatasi pada metode dasar yang umum digunakan untuk meng-ekstrak fitur-fitur dari teks yang dapat digunakan dalam pembelajaran mesin.

a. Bag of Words

   Metode ini menkonversi data dalam bentuk text menjadi vector yang berisi semua kata dari suatu vocabulary tertentu disertai dengan jumlah kemunculan setiap kata pada teks. Contoh penggunaan metode bag-of-words ditunjukkan pada gambar 32, teks dikonversi kedalam vector, dimana kata “it” muncul sebanyak 2 kali, kata “puppy” sebanyak 1 kali, “cute” sekali, dst. Jika kata dalam vocabulary yang digunakan tidak muncul, maka nilainya 0.



Gambar 32. Illustrasi metode bag-of-words (sumber: Zheng & Casari, 2018

Penting untuk diketahui bahwa dalam metode ini, urutan kata dalam vector tidak diperhatikan, yang penting vocabulary (kumpulan kata-kata) yang digunakan konsistent untuk semua teks dalam dataset. Metode ini juga tidak mengenal hirarki diantara kata-kata, semua kata merupakan elemen vector yang setara. Sebagai hasil akhir, jika jumlah kata dalam vocabulary yang digunakan adalah N, maka metode ini akan menghasilkan fitur berukuran N-dimensi.

b. Bag of N-Grams

Metode bag-of-n-grams merupakan ektensi dari metode bag-of-words. Satu n-gram merupakan satu urutan n-token, dimana satu kata adalah 1-gram yang dikenal sebagai *unigram*. Tokenisasi adalah proses untuk memecah teks dalam bagian-bagian yang lebih kecil yang disebut token. Token dapat berupa karakter, kata ataupun kumpulan kata. Dibandingkan metode bag-of-words, metode ini memperhatikan urutan dari kata sehingga menghasilkan fitur yang lebih kaya namun membutuhkan sumber daya komputasi yang lebih besar. Tabel 1 menunjukkan contoh penggunaan n-grams untuk teks “Perubahan iklim menyebabkan hilangnya biodiversitas secara mendadak pada abad ini”.

Tabel 4. Contoh tokenisasi n-grams

n-grams	Token
1-gram	Perubahan, iklim, menyebabkan, hilangnya, biodiversitas, secara, mendadak, pada, abad, ini
2-grams	Perubahan iklim, iklim menyebabkan, menyebabkan hilangnya, hilangnya biodiversitas, biodiversitas secara, secara mendadak, mendadak pada, pada abad, abad ini
3-grams	Perubahan iklim menyebabkan, iklim menyebabkan hilangnya, menyebabkan hilangnya biodiversitas, hilangnya biodiversitas secara, biodiversitas secara mendadak, secara mendadak pada, mendadak pada abad, pada abad ini

Dengan menggunakan token yang dikumpulkan tersebut, vector bag-of-n-

grams dibangun dengan menghitung kemunculan token-token tersebut setiap dokumen dalam dataset.

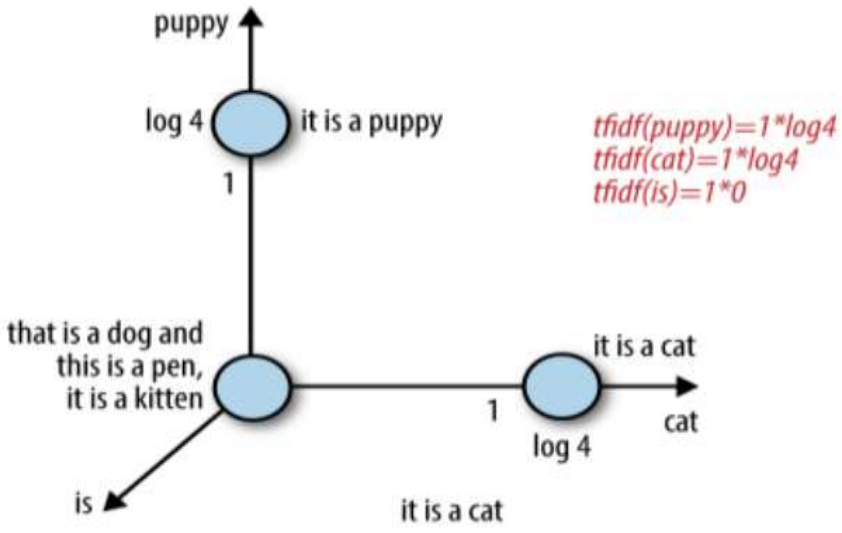
c. TF-IDF

Secara umum metode *Term Frequency* (TF) - *Inverse Document Frequency* (IDF) melakukan normalisasi kemunculan kata (term) dari dokumen dalam dataset, dimana jumlah kemunculan term akan dibagi dengan jumlah dokumen dimana term tersebut muncul. TF mengukur frekuensi kemunculan term dalam suatu dokumen, sedangkan IDF mengukur seberapa penting suatu term. Term yang muncul di banyak dokumen akan mempunyai nilai IDF yang kecil karena dianggap tidak penting. Sebagai contoh, kata sambung seperti “dan”, “di”, “atau” akan muncul hampir disemua dokumen dalam dataset sehingga dapat diabaikan. Secara matematis, TF-IDF dapat dihitung menggunakan rumus:

tf-idf(w, d) = bow(w, d) * log (N / Jumlah dokumen dimana w muncul)

dimana:

- w = word
- d = document
- bow = bag-of-words
- N = jumlah dokumen dalam dataset



Gambar 33. Illustrasi perhitungan TF-IDF (sumber: Zheng & Casari, 2018)

Illustrasi perhitungan TF-IDF ditampilkan pada gambar 6. Pada contoh ini, dataset terdiri dari 4 kalimat yaitu “it is a puppy,” “it is a cat,” “it is a kitten,” dan “that is a dog and this is a pen”. Kalimat tersebut diplot dalam ruang fitur dari 3 kata: “puppy,” “cat,” dan “is.”. Dengan menggunakan rumus TF-IDF diatas, maka diperoleh nilai tf-idf untuk kata “puppy” dan “cat” sekitar 0.6 sedangkan “is” menjadi 0.0 sebab muncul disemua kalimat. Dengan demikian dapat diketahui bahwa kalimat pertama itu lebih dekat ke fitur “puppy”, kalimat kedua ke fitur “cat” dan kalimat 3 & 4 tidak dekat ke salah satu fitur termasuk “is” yang mempunyai nilai tf-idf = 0.

C. Seleksi Fitur

Seleksi fitur (*feature selection*) merupakan proses untuk menseleksi fitur-fitur terbaik yang dapat memberikan hasil pembelajaran mesin yang lebih baik. Jika dataset mempunyai N-fitur, proses seleksi fitur adalah mencari subset dari N yang dapat memperbaiki proses pembelajaran yang dilakukan, misalnya menghilangkan noise dari dataset. Berbeda dengan noise yang dapat dihilangkan menggunakan metode prapemrosesan data seperti standarisasi atau normalisasi sebagaimana dijabarkan pada materi sebelumnya, noise pada fitur mengacu pada kemampuan fitur untuk memberikan hasil yang baik dalam proses pembelajaran mesin. Semakin kurang hasil yang diberikan, maka noise yang terdapat dalam fitur tersebut semakin tinggi sehingga semakin tidak relevan untuk digunakan.

Penentuan ukuran hasil pembelajaran mesin tergantung pada tujuan dan metode pembelajaran yang akan digunakan. Sebagai contoh, akurasi [18] umumnya digunakan sebagai ukuran dalam proses klasifikasi, sedangkan untuk regresi umumnya digunakan *Root Mean Square Error* (RMSE)[19]. Selain itu dapat juga digunakan True dan False Positive rate, termasuk False Negative rate[20]. Selain itu, dapat juga diukur hal yang tidak berkaitan langsung dengan pembelajaran yang dilakukan seperti waktu yang digunakan untuk melakukan training, jumlah item data latih yang dibutuhkan untuk memperoleh hasil yang optimal, dan

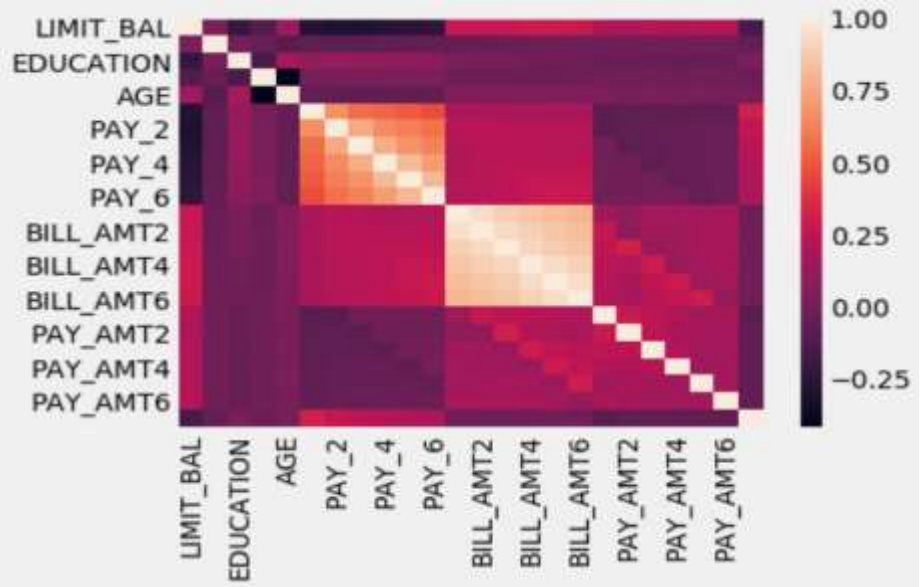
[18] https://en.wikipedia.org/wiki/Accuracy_and_precision
[19] https://en.wikipedia.org/wiki/Root-mean-square_deviation
[20] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

sebagainya.

Secara umum, teknik seleksi fitur terbagi atas tiga, yaitu filtering, wrapper, dan embedded (Zheng & Casari, 2018). Ketiga teknik tersebut akan dijelaskan pada bagian berikut ini.

1. **Filtering**

   Teknik ini digunakan untuk menghilangkan fitur yang tidak relevan dengan hasil yang diinginkan, misalnya akan menghasilkan akurasi yang rendah. Teknik filtering yang umum digunakan adalah menghitung korelasi antara setiap fitur dengan variabel output lalu menghapus fitur dengan yang mempunyai korelasi rendah (nilai korelasi dibawah threshold tertentu). Koefisien Pearson Correlation[21] banyak digunakan untuk menghitung korelasi linier antara 2 set data. Gambar 7 menampilkan contoh korelasi antara fitur menggunakan heatmap[22]. Dalam hal ini, nilai koefisien adalah antara -1 dan +1 dimana nilai 0 menandakan tidak korelasi sama sekali sedangkan nilai yang mendekati -1 dan +1 menandakan korelasi yang sangat kuat. Dengan menetapkan nilai threshold tertentu (misal 0.2), maka fitur yang mempunyai nilai korelasi dibawah nilai tersebut akan tidak dijadikan fitur dalam pembelajaran mesin.



Gambar 34. Heatmap yang menggambarkan korelasi antar variabel (sumber: Ozdemir & Susarla, 2018)

[21] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

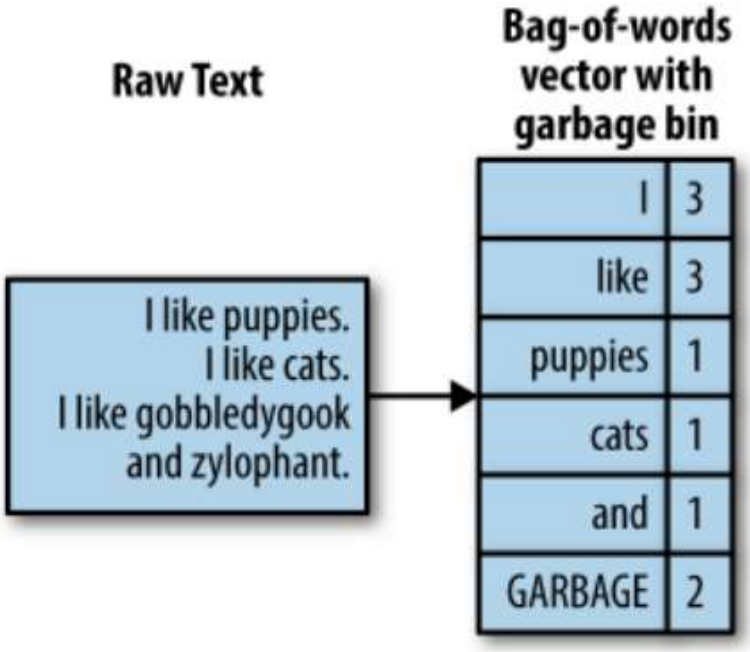
[22] https://en.wikipedia.org/wiki/Heat_map

Teknik filtering juga dapat digunakan untuk data yang berbentuk teks misalnya dengan menghilangkan kata yang tidak berhubungan sama sekali, seperti kata penutup (stopwords), kata umum ataupun kata yang sangat tidak umum. Gambar 8 menunjukkan contoh kata yang sering digunakan dalam review Yelp, dimana kata yang paling sering digunakan itu berupa kata bantu atau sambung “the”, “and”, “a”, dst. Terlihat bahwa kata-kata tersebut sangat umum dan tidak terkait dengan analisis yang akan dilakukan. Dengan menyaring kata-kata tersebut maka akan dihasilkan dataset yang terdiri dari kata-kata yang relavan dengan tujuan analis data yang akan dilakukan.

Rank	Word	Document frequency	Rank	Word	Document frequency
1	the	1416058	21	t	684049
2	and	1381324	22	not	649824
3	a	1263126	23	s	626764
4	i	1230214	24	had	620284
5	to	1196238	25	so	608061
6	it	1027835	26	place	601918
7	of	1025638	27	good	598393
8	for	993430	28	at	596317
9	is	988547	29	are	585548
10	in	961518	30	food	562332
11	was	929703	31	be	543588
12	this	844824	32	we	537133
13	but	822313	33	great	520634
14	my	786595	34	were	516685
15	that	777045	35	there	510897
16	with	775044	36	here	481542
17	on	735419	37	all	478490
18	they	720994	38	if	475175
19	you	701015	39	very	460796
20	have	692749	40	out	460452

Gambar 35. Kata yang paling sering digunakan dalam dataset review Yelp (sumber: Zheng & Casari, 2018).

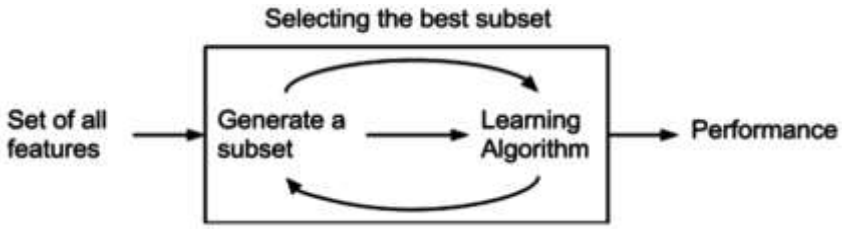
Contoh lainnya adalah satu kategori khusus untuk menampung kata yang tidak dikenali atau tidak umum. Gambar 9 menunjukkan contoh penggunaan kategori khusus tersebut, dimana dari teks ditemukan 2 kata yang tidak umum yaitu “gobbledygook” dan “zylophant”. Kedua kata tersebut selanjutnya difilter dengan dimasukkan dalam 1 kategori khusus GARBAGE.



Gambar 36. Contoh penyediaan kategori untuk menampung kata-kata yang tidak umum digunakan (sumber: Zheng & Casari, 2018)

2. Wrapper

Kekurangan metode filtering yang dijelaskan sebelumnya adalah tidak memperhitungkan kontribusi fitur secara komulatif. Misalnya, ada fitur yang jika digunakan secara individual maka tidak mempunyai pengaruh pada hasil, tetapi akan berpengaruh jika digunan secara bersama-sama dengan fitur yang lain. Teknik Wrapper mengatasi kekurangan tersebut dengan mencari kombinasi (subset) fitur yang menghasilkan hasil yang terbaik.

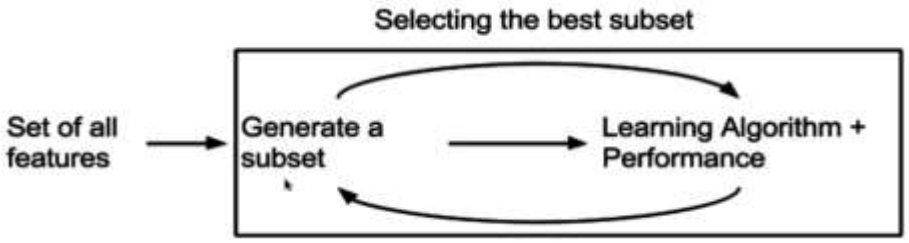


Gambar 37. Mekanisme seleks fitur menggunakan teknik Wrapper (sumber: Wikipedia.org)

Gambar 37 menunjukkan mekanisme seleksi fitur menggunakan teknik Wrapper. Pertama, sejumlah fitur akan dipilih menjadi kandidat subset fitur. Kedua, kandidat tersebut digunakan untuk melakukan pembelajaran mesin, dimana tingkat keberhasilan output dicatat. Proses tersebut diulang-ulang sampai diperoleh kandidat subset fitur yang menghasilkan output yang terbaik.

3. Embedded

Teknik ini menggabungkan kelebihan dari kedua teknik sebelumnya, dimana seleksi fitur dilakukan secara bersamaan proses pembelajaran mesin. Pada teknik ini, algoritma pembelajaran mesin akan menyesuaikan paramater internal dan menghitung bobot dari setiap fitur yang menunjukkan tingkat pengaruh dari setiap fitur terhadap output (Pudjihartono et al., 2022). Gambar 11 menunjukkan mekansime embedded ini, dimana subset fitur yang terbaik akan ditentukan saat pembelajaran mesin sedang berlangsung.



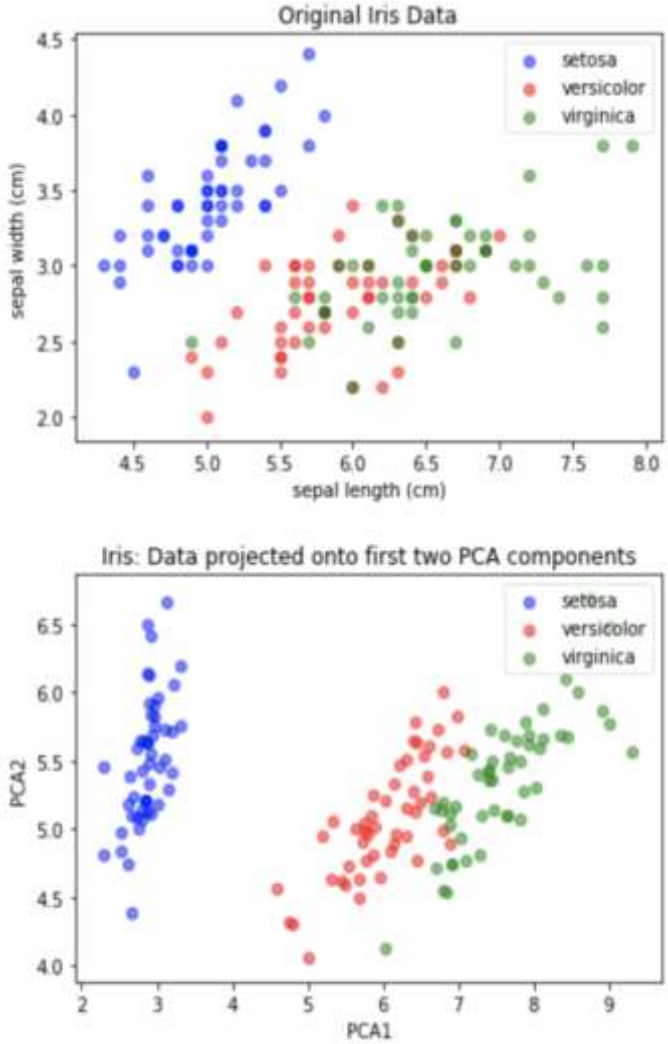
Gambar 38. Mekanisme seleksi fitur menggunakan teknik Embedded (sumber: Wikipedia.org)

D. Transformasi Fitur

Keluaran dari metode seleksi fitur yang dijelaskan pada bagian sebelumnya adalah berupa kumpulan (subset) fitur yang dianggap terbaik untuk melakukan pembelajaran mesin. Ada kalanya dibutuhkan proses untuk mentransformasi fitur-fitur yang menjadi fitur yang baru yang lebih baik. Sebagai contoh, jika dimensi data kita terlalu besar, maka pembelajaran yang dilakukan membutuhkan waktu yang lama dengan hasil yang tidak memuaskan, biasa juga disebut curse of dimensionality[23]. Jika pada pengembangan fitur yang dijelaskan sebelumnya, fitur baru dihasilkan dengan melakukan operasi (misal

[23] https://en.wikipedia.org/wiki/Curse_of_dimensionality

penjumlahan, perkalian) pada fitur-fitur yang sudah ada. Sedangkan transformasi fitur pada dasarnya mengambil informasi dari semua fitur untuk menghasilkan representasi baru dari fitur-fitur tersebut. Algorithma yang umum digunakan adalah *Principal Component Analysis* (PCA)[24] dan *Linear Discriminant Analysis* (LDA)[25]. Secara umum, algoritma tersebut melakukan proyeksi item-item data ke koordinat baru (di PCA disebut principal components) pada dimensi yang lebih rendah. PCA fokus pada variasi data secara keseluruhan, sedangkan LDA fokus pada fitur yang mempunyai mempunyai kemampuan pembeda yang terbaik. Gambar 12 menunjukkan penggunaan PCA untuk meproyeksikan dataset Iris[26] ke 2 principal komponen pertama.



Gambar 39. Contoh transformasi fitur dari data dengan dimensi tinggi ke data dimensi rendah (sumber: Ozdemir & Susarla, 2018)

[24] https://en.wikipedia.org/wiki/Principal_component_analysis
[25] https://en.wikipedia.org/wiki/Linear_discriminant_analysis
[26] https://en.wikipedia.org/wiki/Iris_flower_data_set

E. Rangkuman

Perekayasaan fitur merupakan metode prapemrosesan data yang dimaksudkan untuk mentransformasi data ke dalam bentuk yang lebih sesuai untuk digunakan dalam proses pembelajaran mesin. Berbeda dengan metode prapemrosesan data pada umumnya, metode ini bekerja pada level fitur yaitu attribut data dalam kaitannya dengan tujuan pembelajaran mesin yang akan dilakukan. Perekayasaan fitur dapat dilakukan dengan mentransformasi data kategorikal kedalam bentuk data numerikal, data tekstual kedalam bentuk data numerikal. Dapat juga dilakukan seleksi fitur yang memberikan hasil pembelajaran yang terbaik menggunakan berbagai teknik seperti filtering, wrapper, maupun embedded. Sebagai catatan terakhir, perekayasaan fitur tidak bisa berdiri sendiri, dia harus dikaitkan dengan tujuan pembelajaran yang akan dilakukan. Jika ada 2 kegiatan dengan tujuan berbeda, walaupun menggunakan dataset yang sama, maka perekayasaan fitur yang dibutuhkan juga berbeda. Lebih dari itu, perekayasaan fitur juga dipengaruhi oleh bentuk data itu sendiri.

Tabel berikut ini menunjukkan beberapa referensi online terkait alat bantu untuk melakukan perekayasaan fitur.

No	Keterangan	Link
1	One Hot Encoding	https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
2	Feature Extraction	https://scikit-learn.org/stable/modules/feature_extraction.html
3	Feature Selection	https://scikit-learn.org/stable/modules/feature_selection.html
4	Principal Component Analysis	https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
5	Linear Discriminant Analysis	https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

F. Evaluasi

1. Identifikasi dataset dari kegiatan dalam unit kerja atau instansi anda yang membutuhkan perekayasaan fitur. Misalnya, lakukan identifikasi:
   a. Apakah datanya berbentuk tekstual
   b. Apakah datanya berbentuk kategorikal
2. Berikan penjelasan terkait fitur-fitur dari dataset yang anda identifikasi pada nomor 1, seperti:
   a. Struktur data
   b. Kategori data
   c. Level data
3. Gunakan pengembangan fitur yang sesuai dengan fitur yang anda identifikasi pada nomor 2.
4. Tentukan subset fitur yang paling baik untuk fitur yang anda kembangkan dari nomor 3 dengan menggunakan teknik seleksi fitur yang telah anda pelajari.

DAFTAR PUSTAKA

Akbar, Z., Wardani, W., Mahendra, T., Kartika, Y. A., Indrawati, A., Djarwaningsih, T., Manik, L. P., & Yaman, A. (2022). Semantic Annotation of Objects of Interest in Digitized Herbarium Specimens for Fine-Grained Object Classification. In A. Patel, N. C. Debnath, & B. Bhushan, Semantic Web Technologies (1st ed., pp. 181–202). CRC Press. https://doi.org/10.1201/9781003309420-8

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Consoli, S., Barbaglia, L., & Manzan, S. (2021). Explaining Sentiment from Lexicon. Joint Proceedings of the 2nd International Workshop on Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP 2021) & 6th International Workshop on Explainable Sentiment Mining and Emotion Detection (X-SENTIMENT 2021) co-located with co-located with 18th Extended Semantic Web Conference 2021. https://ceur-ws.org/Vol-2918/paper7.pdf

Dutta, A., & Zisserman, A. (2019). The VIA Annotation Software for Images, Audio and Video. Proceedings of the 27th ACM International Conference on Multimedia, 2276–2279. https://doi.org/10.1145/3343031.3350535

Fernandes, J.D., Hinrichs, A.S., Clawson, H. et al. The UCSC SARS-CoV-2 Genome Browser. Nat Genet 52, 991–998 (2020). https://doi.org/10.1038/s41588-020-0700-8

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.-Y., Danielsen, F., Hitchcock, C. B., Hulbert, J. M., Piera, J., Spiers, H., Thiel, M., & Haklay, M. (2022). Citizen science in environmental and ecological sciences. Nature Reviews Methods

Primers, 2(1), 64. https://doi.org/10.1038/s43586-022-00144-4
Gura, T. (2013). Citizen science: Amateur experts. Nature, 496(7444), 259–261. https://doi.org/10.1038/nj7444-259a

Hardison, D. R., Holland, W. C., Currier, R. D., Kirkpatrick, B., Stumpf, R., Fanara, T., Burris, D., Reich, A., Kirkpatrick, G. J., & Litaker, R. W. (2019). HABscope: A tool for use by citizen scientists to facilitate early warning of respiratory irritation caused by toxic blooms of Karenia brevis. PLOS ONE, 14(6), e0218489. https://doi.org/10.1371/journal.pone.0218489

Hernández, M. A., & Stolfo, S. J. (1998, January). Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2(1), 9–37.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., & Zapata, F. (2005). Plant Ontology (PO): A Controlled Vocabulary of Plant Structures and Growth Stages. Comparative and Functional Genomics, 6(7–8), 388–397. https://doi.org/10.1002/cfg.496

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. Data Mining and Knowledge Discovery, 7(1), 81–99.
Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. Journal of Web Semantics, 2(1), 49–79. https://doi.org/10.1016/j.websem.2004.07.005

Kohler, K. E., & Gill, S. M. (2006). Coral Point Count with Excel extensions (CPCe): A Visual Basic program for the determination of coral and substrate coverage using random point count methodology. Computers & Geosciences, 32(9), 1259–1269. https://doi.org/10.1016/j.cageo.2005.11.009

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), 111–117.
Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts*

and practice with rapidminer. Morgan Kaufmann.

Loshin, D. (2012). Business intelligence: the savvy manager's guide. Newnes.
MacDougall, A., Volynkin, V., Saidi, R., Poggioli, D., Zellner, H., Hatton-Ellis, E., Joshi, V., O'Donovan, C., Orchard, S., Auchincloss, A. H., Baratin, D., Bolleman, J., Coudert, E., de Castro, E., Hulo, C., Masson, P., Pedruzzi, I., Rivoire, C., Arighi, C., … The UniProt Consortium. (2020). UniRule: A unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics, btaa485. https://doi.org/10.1093/bioinformatics/btaa485

Ozdemir, S., & Susarla, D. (2018). Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems. Packt Publishing Ltd.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics, 2, 927312
Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. The VLDB Journal, 29(2–3), 709–730. https://doi.org/10.1007/s00778-019-00552-1

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., & Ré, C. (2019). Training Complex Models with Multi-Task Weak Supervision. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 4763–4771. https://doi.org/10.1609/aaai.v33i01.33014763

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, 77(1–3), 157–173. https://doi.org/10.1007/s11263-007-0090-8

Sharda, R., Delen, D., & Turban, E. (2017). *Business intelligence, analytics, and data science: a managerial perspective*. pearson.
Vercellis, C. (2011). *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.

Walls, R. L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M. A., Jaiswal, P., Mungall, C. J., Preece, J., Rensing, S., Smith, B., & Stevenson, D. W. (2012). Ontologies as integrative tools for plant science. American Journal of Botany, 99(8), 1263–1275. https://doi.org/10.3732/ajb.1200222

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc.