

Menganalisa Konten Percakapan di Media Sosial



HI!

Muhammad Apriandito Arya Saputra

- Mahasiswa MBA-ITB
- Data Scientist, sejak 2017
- Pengurus Asosiasi Ilmuwan Data Indonesia
- CEO Technaut Education

Email : muhammad-apriandito@sbm-itb.ac.id

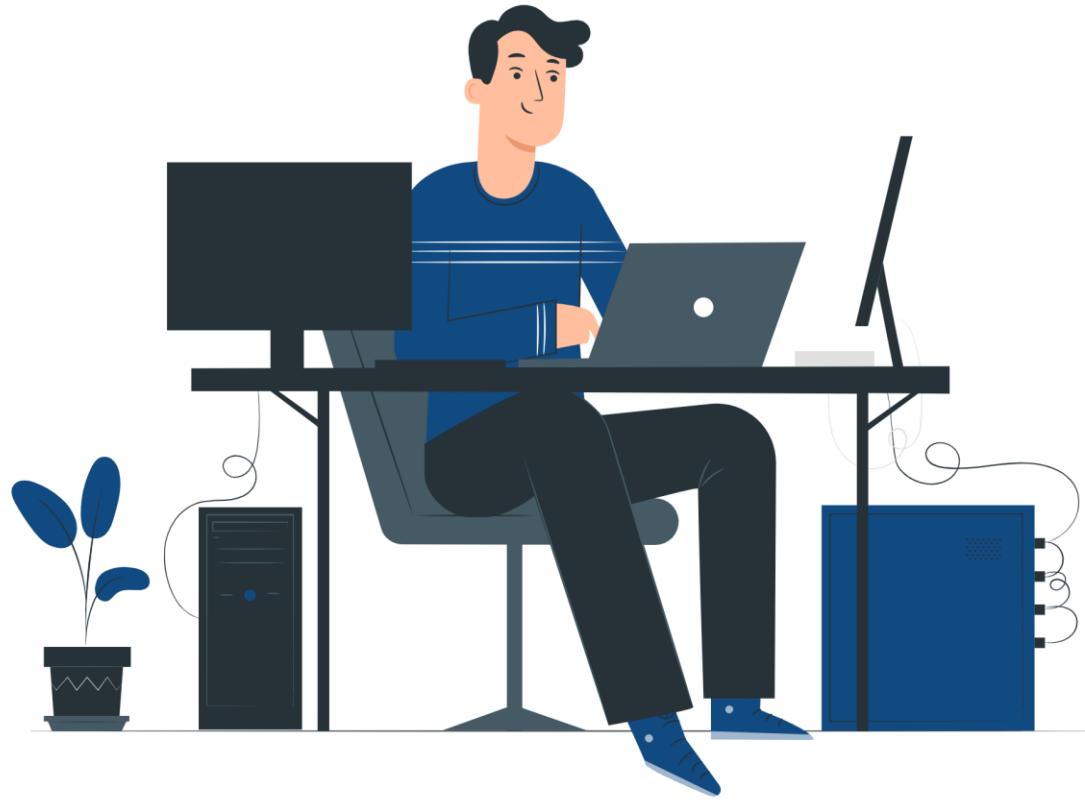
Website : www.apriandito.com

Medium/GitHub/Instagram/SpeakerDeck/Kaggle: [@apriandito](https://www.twitter.com/@apriandito)



Apa saja yang akan kita pelajari?

1. Fenomena Big Data
2. Konsep Analisis Media Sosial
3. Mengambil Data Media Sosial (Twitter)
4. Text Mining
5. Text Classification (Sentiment Analysis)
6. Topic Modelling (*Opsional*)



JAN
2020

DIGITAL AROUND THE WORLD IN 2020

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

TOTAL
POPULATION



7.75
BILLION

URBANISATION:

55%

UNIQUE MOBILE
PHONE USERS



5.19
BILLION

PENETRATION:

67%

INTERNET
USERS



4.54
BILLION

PENETRATION:

59%

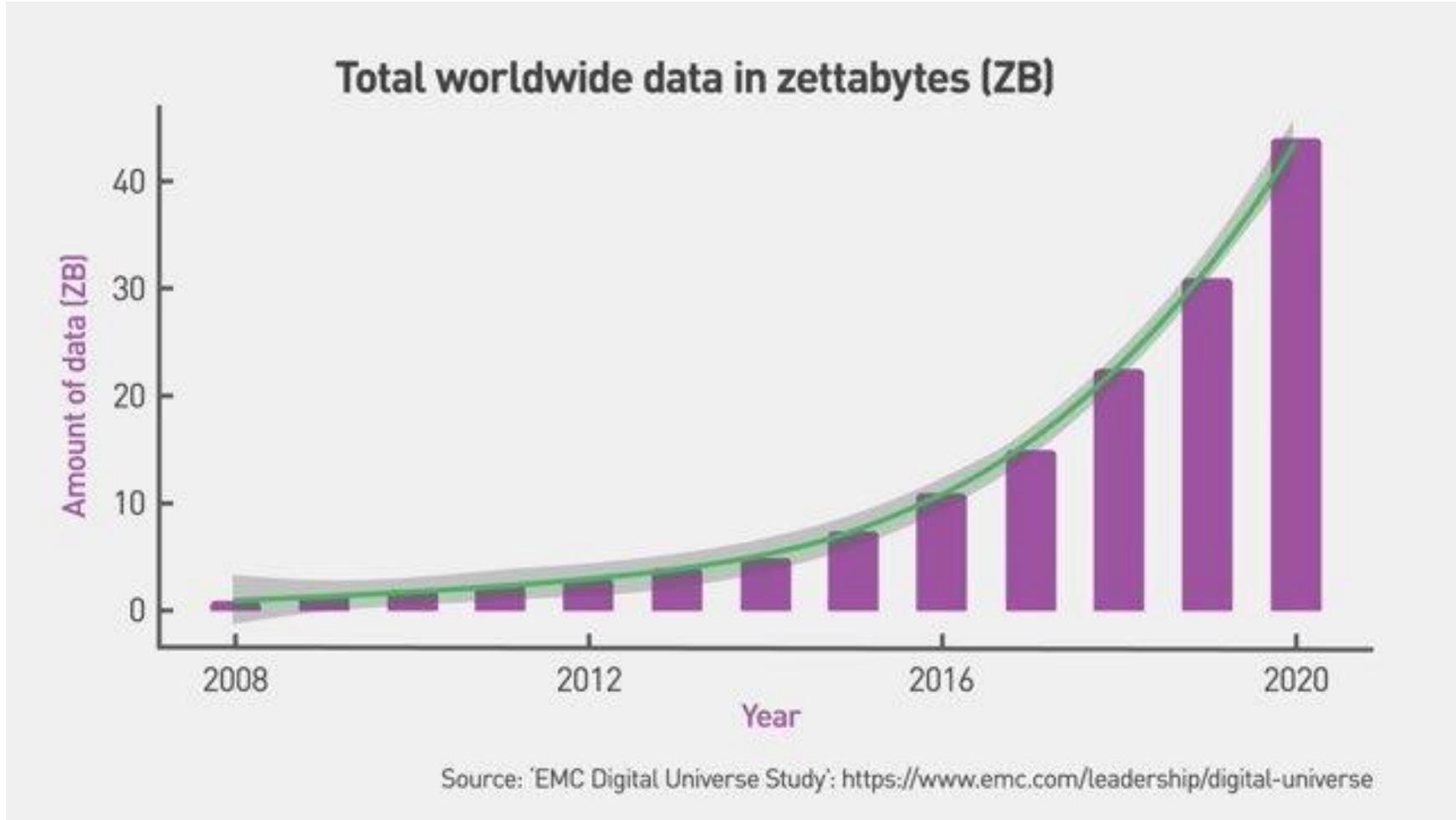
ACTIVE SOCIAL
MEDIA USERS



3.80
BILLION

PENETRATION:

49%

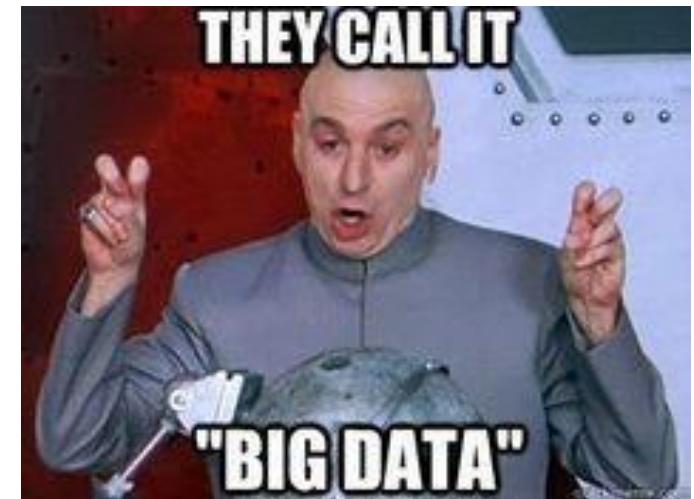


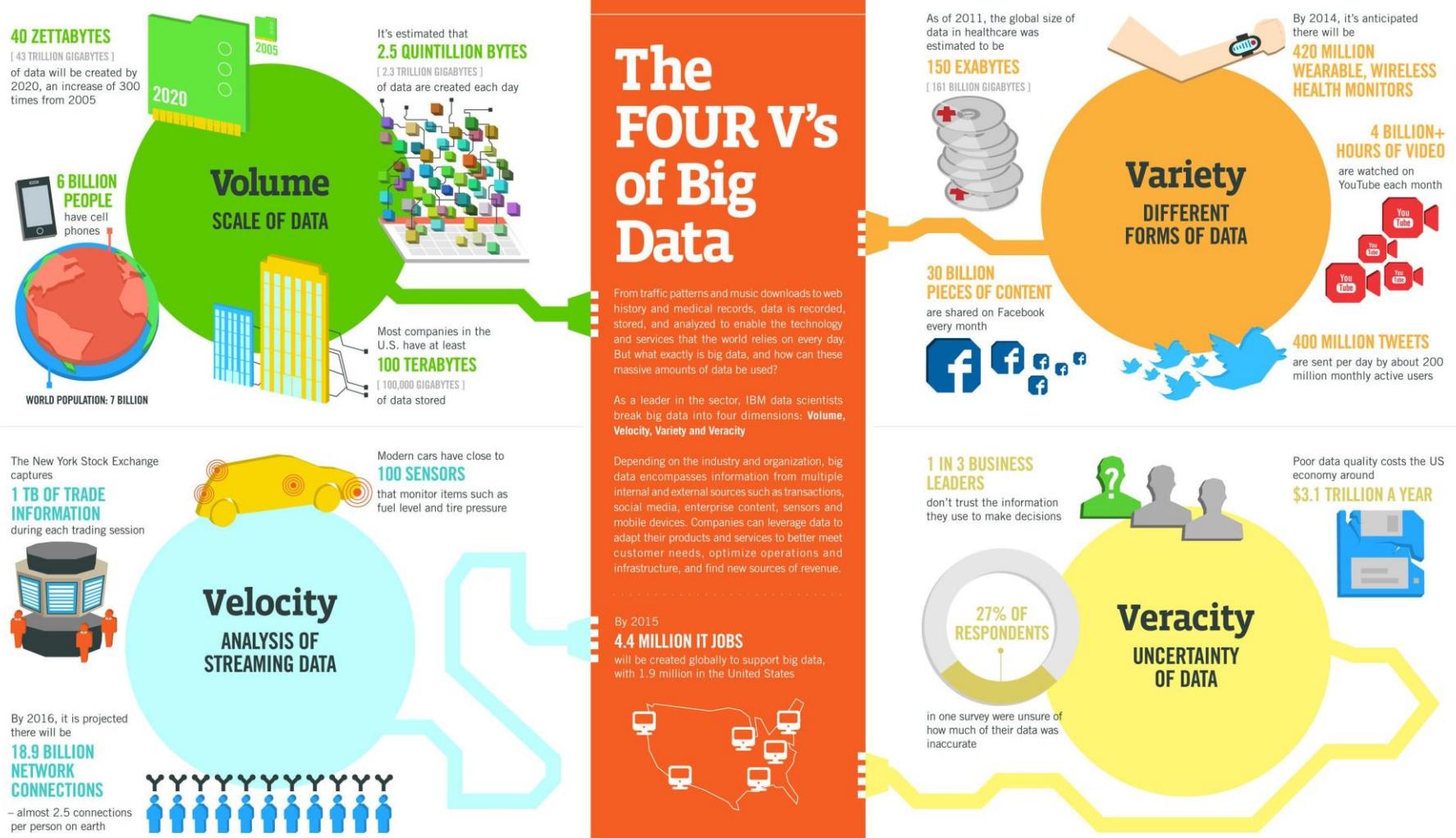
Big Data

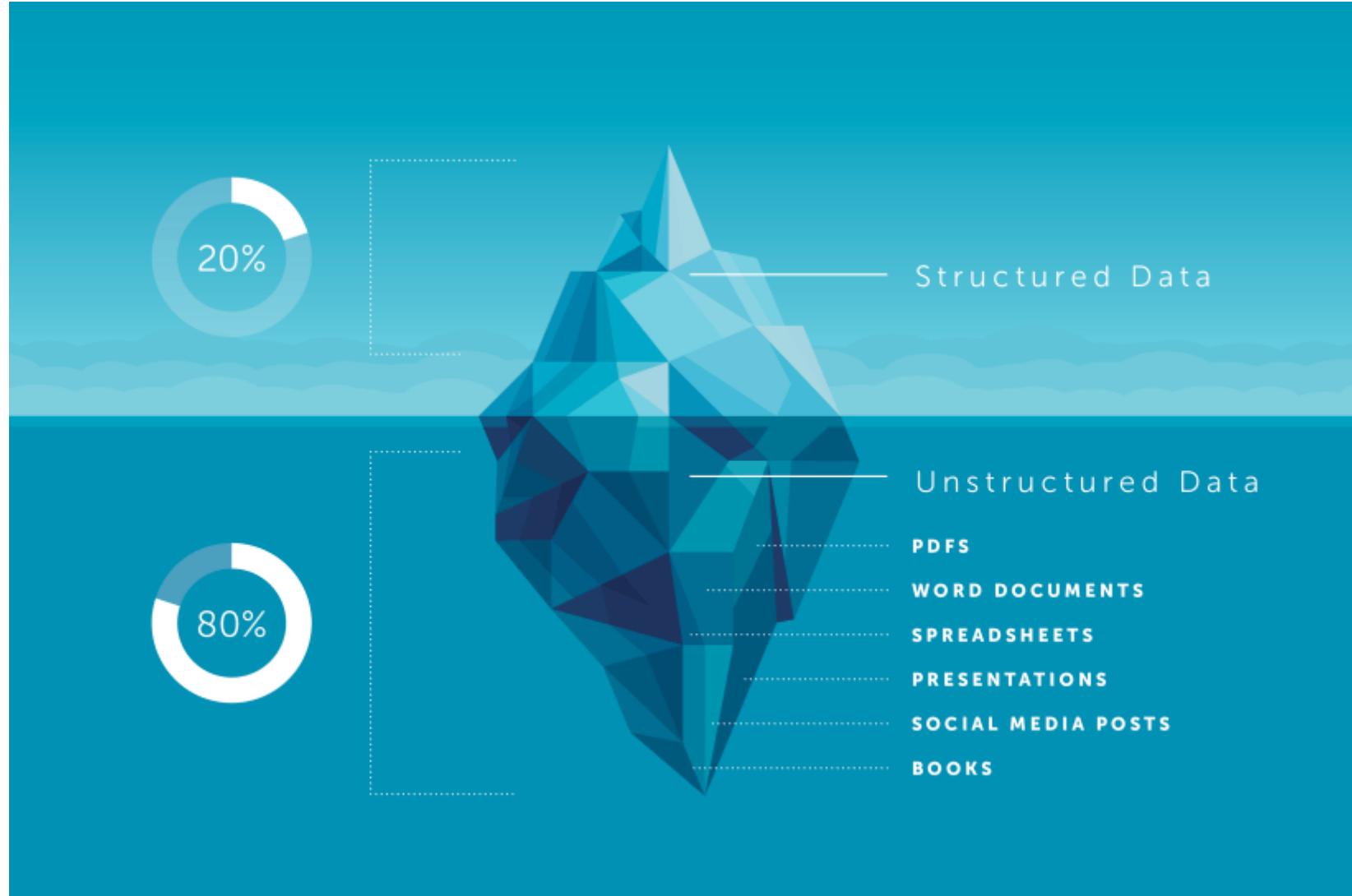
Big Data adalah istilah yang menggambarkan volume data yang besar, baik data yang terstruktur maupun data yang tidak terstruktur.

Big data memiliki 4 Karateristik:

- Volume
- Variety
- Velocity
- Veracity

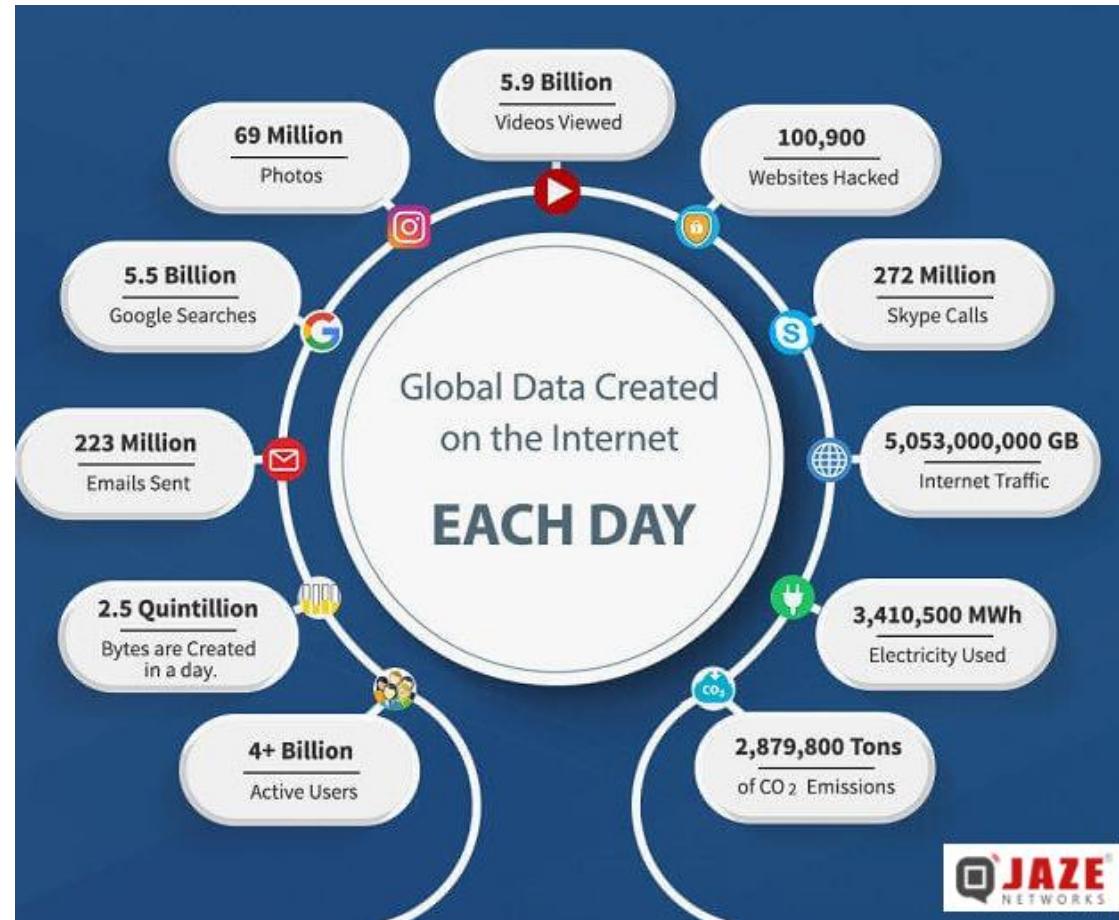






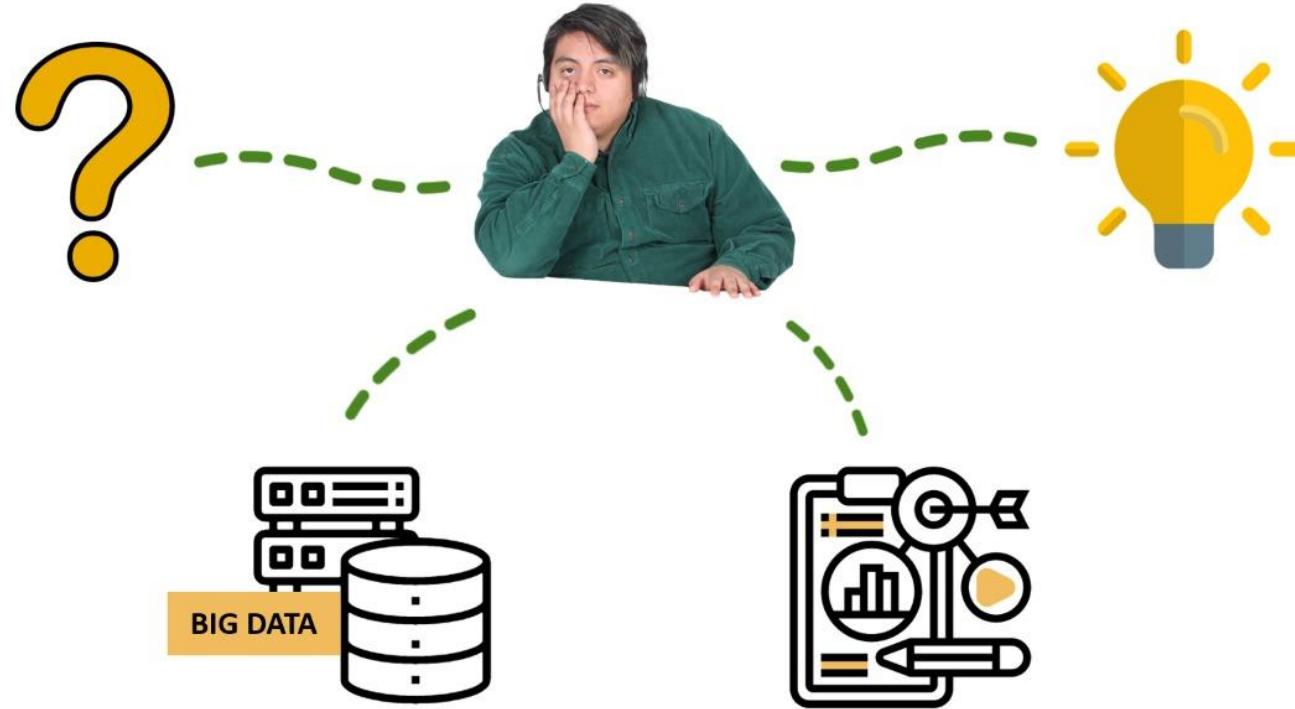
Sumber: <https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/>

Data Social Media



Sumber: www.jazenetworks.com

Creating Value User Behavior Analytics



Foursquare Check In



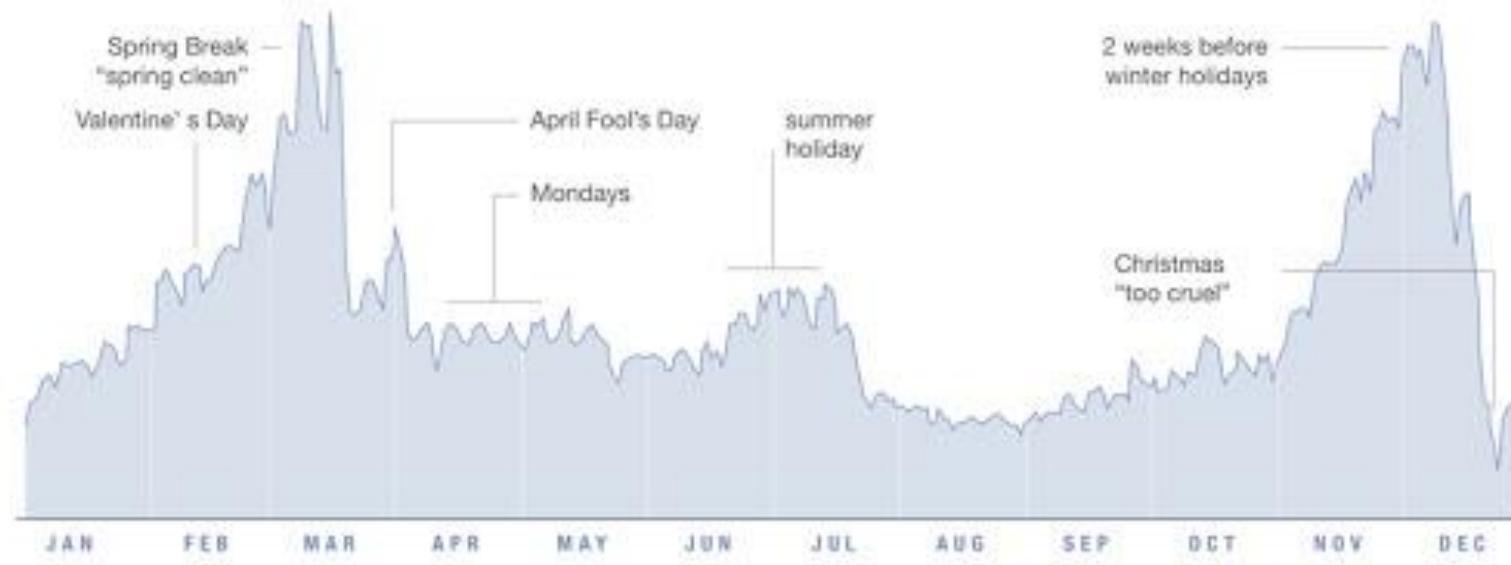
<https://foursquare.com/infographics/pulse>

Tweet Terkait Trump



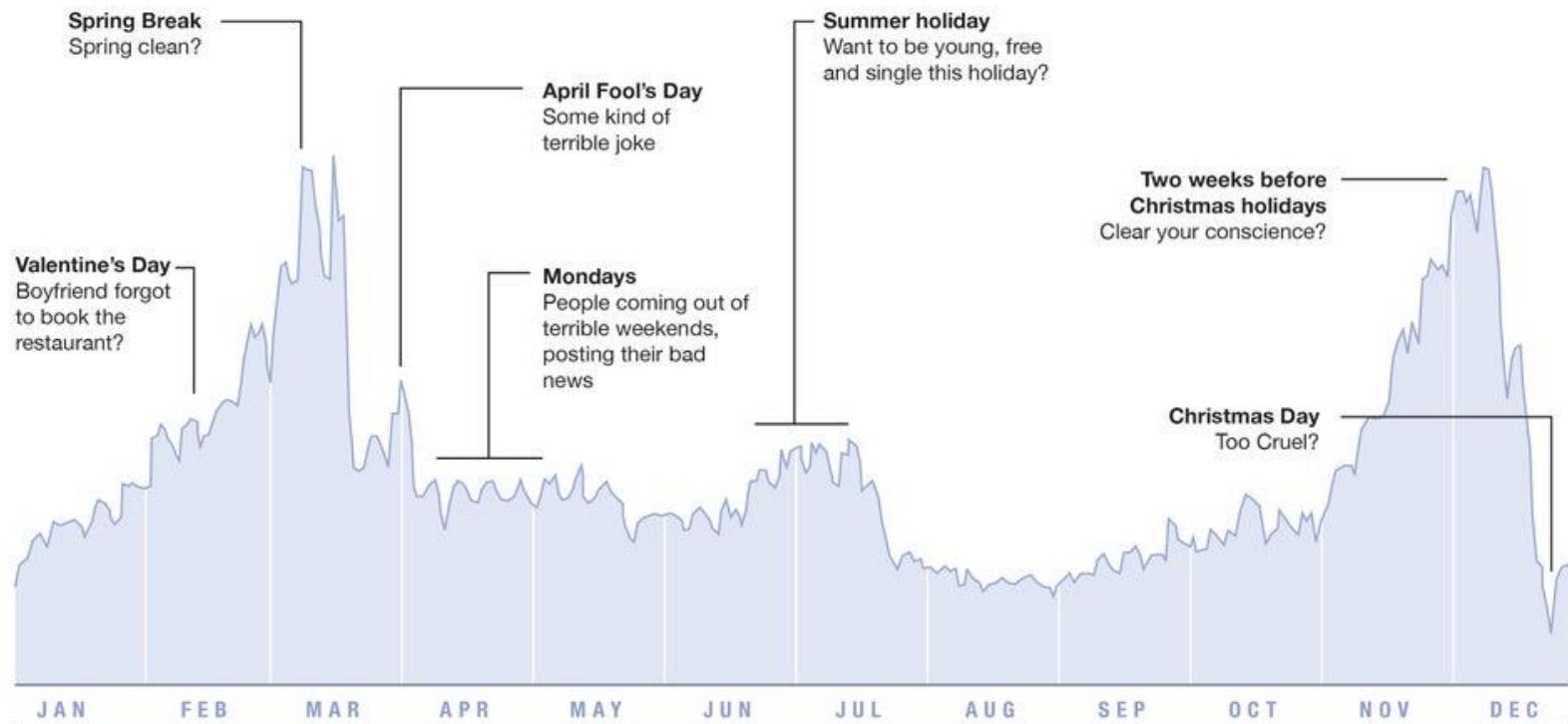
Sumber: <https://www.techrepublic.com/article/a-data-visualization-of-trump-trends-on-social-media/>

Pattern apa ini?



Peak Break-up Times

According to Facebook status updates



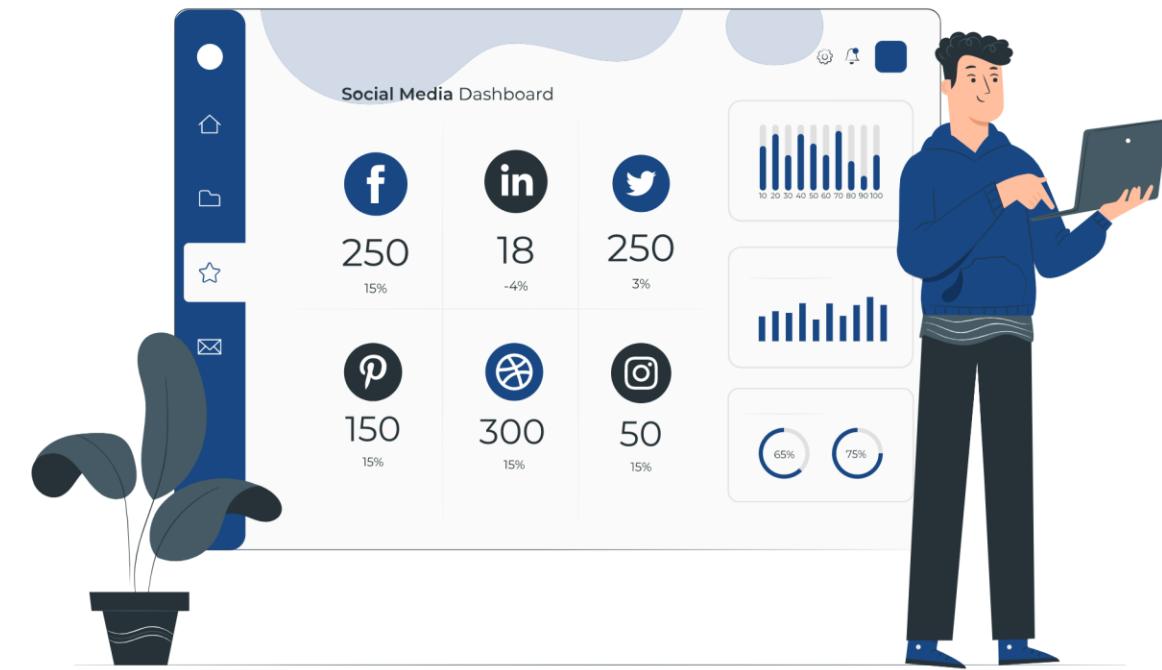
David McCandless & Lee Byron
InformationIsBeautiful.net / LeeByron.com

source: Facebook Lexicon 2008

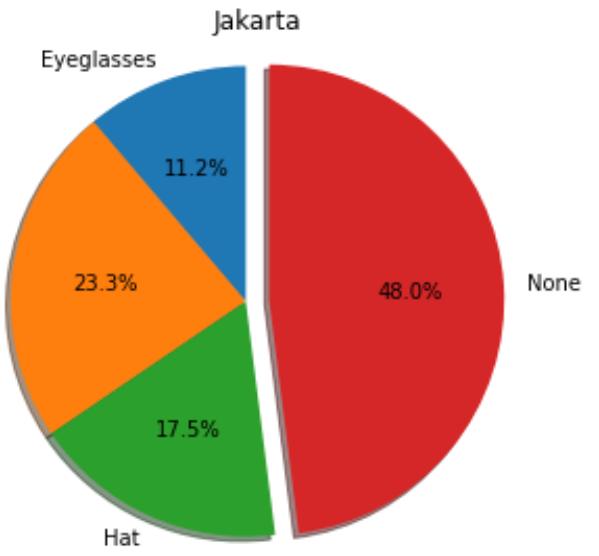
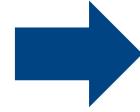
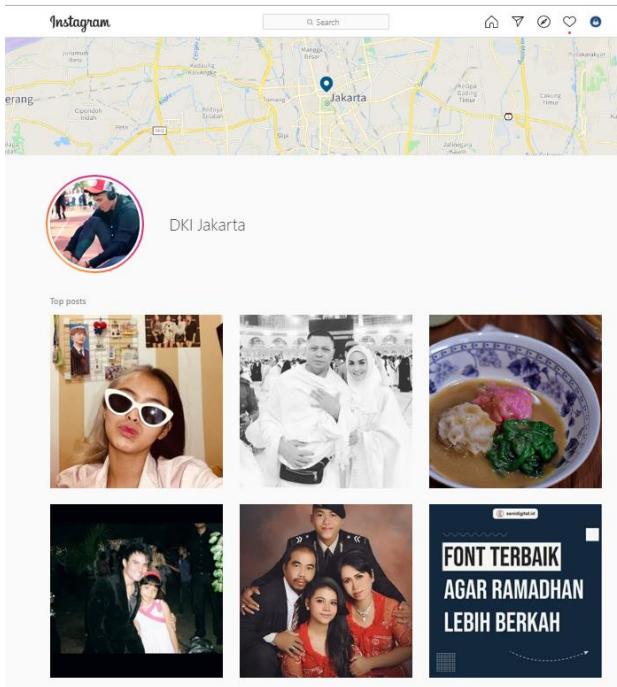
Social Media Analytics

Social Media Analytics (SMA) refers to the approach of collecting data from social media sites and blogs and evaluating that data to make business decisions.

Sumber: Wikipedia



Social Media Analytics: Instagram



Social Media Analytics: Tripadvisor



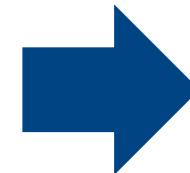
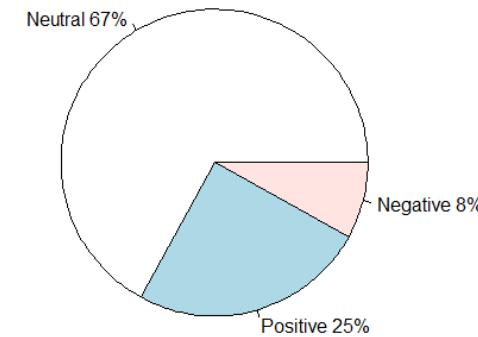
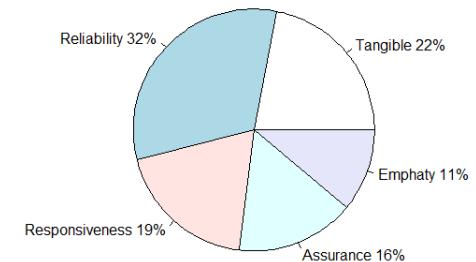
Tripadvisor.

**Bisnis trip**

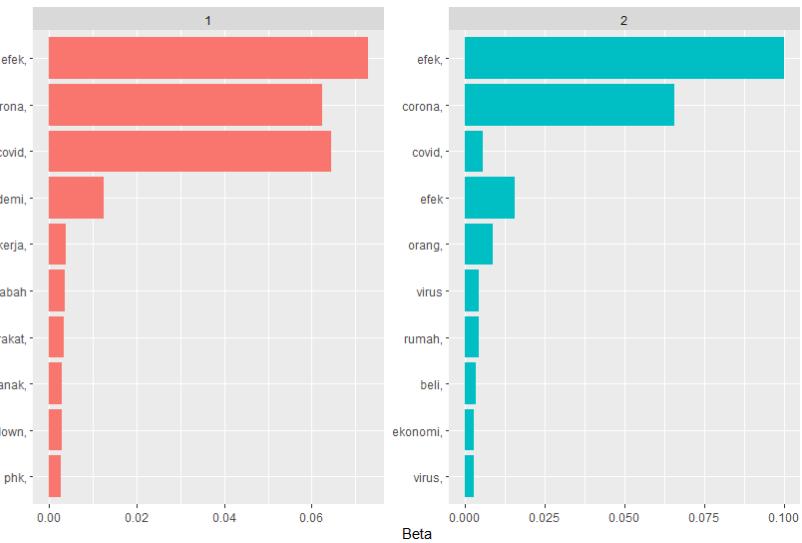
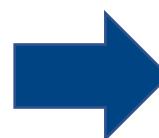
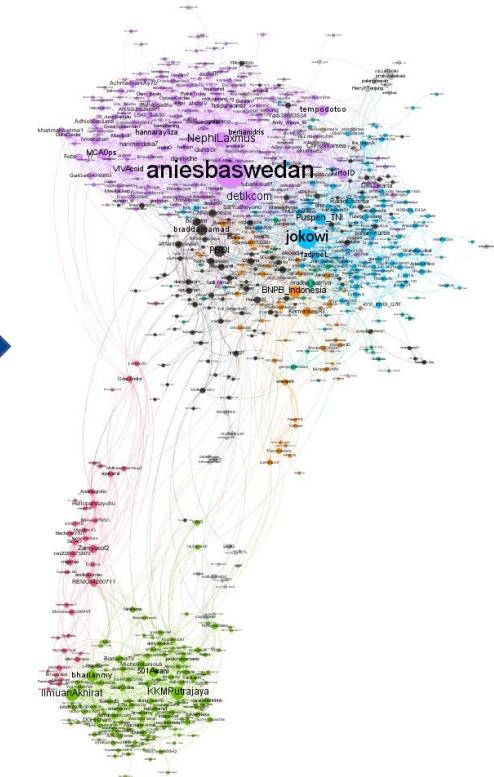
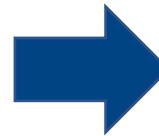
"Hotel dengan kamar yang sangat nyaman, breakfast lengkap dan enak. Disediakan meja kerja cocok untuk yang bisnis trip,next trip akan ajak teman dan keluarga menginap kesini lagi jika berkunjung ke bali."

[Selengkapnya ▾](#)**Lebih dari harga dengan Pelayanan Terburuk**

"Saya menjalankan perusahaan yang berinvestasi dalam keadilan pribadi hotel, resor dan properti, yang memberi saya banyak pengalaman ribuan hotel yang berbeda di seluruh dunia. Tapi aku menyesal untuk mengatakan bahwa Katamama disajikan saya dengan pengalaman terburuk di akomodasi dengan harga tertinggi yang pernah saya alami dalam tahun. Saya menginap di suite di atap untuk melihat dan meningkatkan kemampuan mereka di atas layanan dan fasilitas. Saya dapat mengatakan saya pengalaman terburuk di bawah tiga hal: 1. Fasilitas Berbahaya 2. Lingkungan yang keras 3. Pelayanan Tidak Profesional Pertama, fasilitas sangat modern yang terlihat sedang dibangun oleh grup F & menjadi populer di Potato Head. Namun ketika Anda melihat mereka secara detail, Anda akan

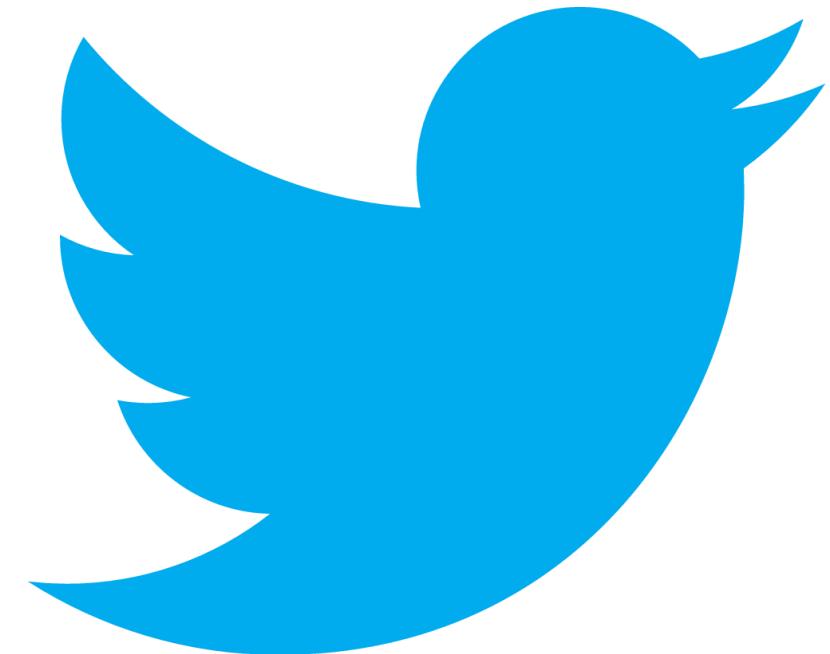
[Selengkapnya ▾](#)**Sentiment Analysis****Service Quality Dimension - Negative**

Social Media Analytics: Twitter



Twitter

Twitter merupakan layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter akan tetapi pada tanggal 07 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebutan kicauan (tweet).



Kelebihan dan Kekurangan

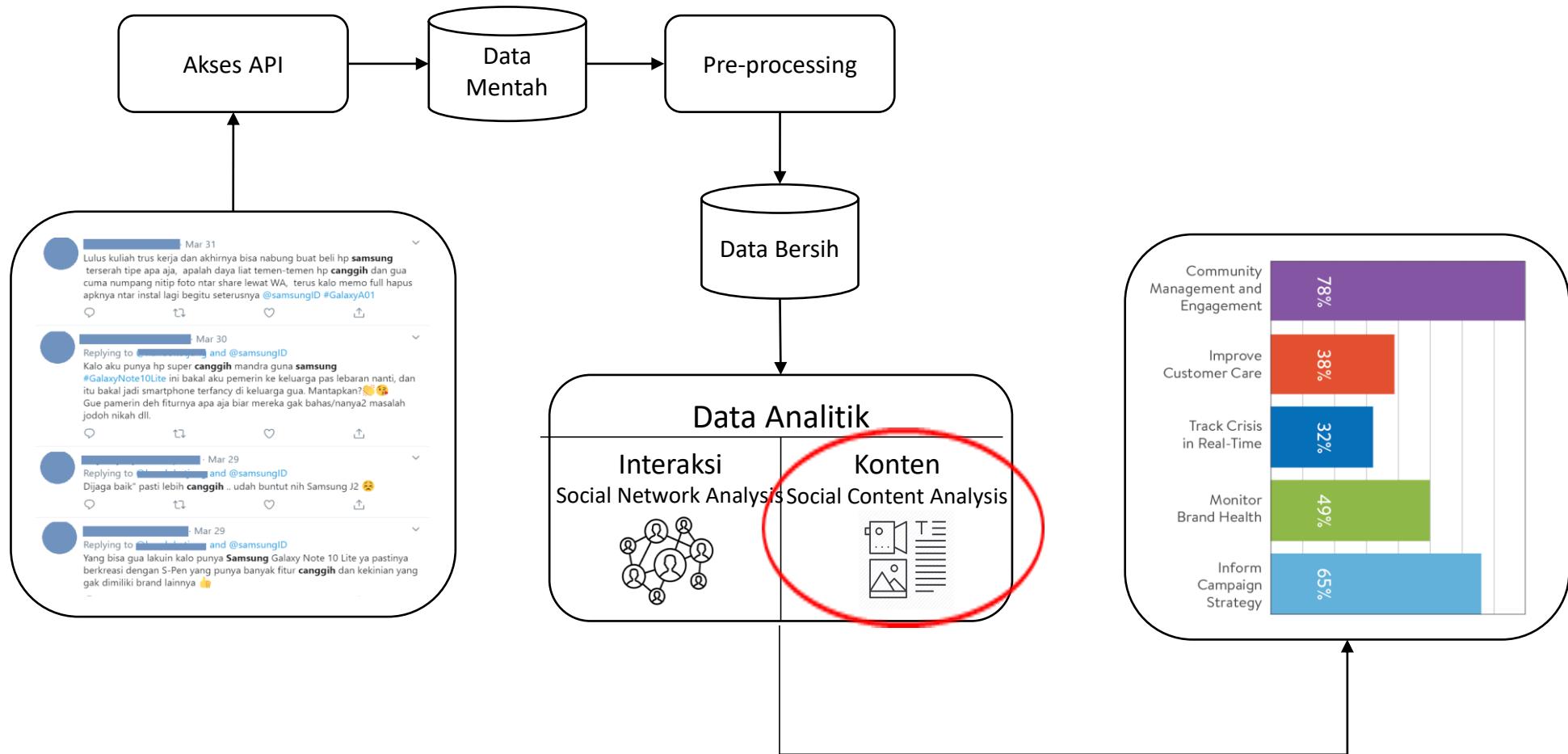
Kelebihan	Kekurangan
<ul style="list-style-type: none">• TwitterAPI terbuka dan dapat diakses• Mudah menemukan percakapan karena adanya fitur Hashtag• Lebih mudah dikontrol dan dianalisa, karena jumlah tweet terbatas.	<ul style="list-style-type: none">• Adanya batasan pengambilan data untuk akun gratis• Tweet yang memiliki penandaan geografis sangat sedikit (1%)

Penggunaan Data Twitter

- Mengidentifikasi trending topik
- Menganalisa opini pelanggan
- Menganalisa sentiment public
- Melihat capaian sebuah merek
- Deteksi Event



Framework



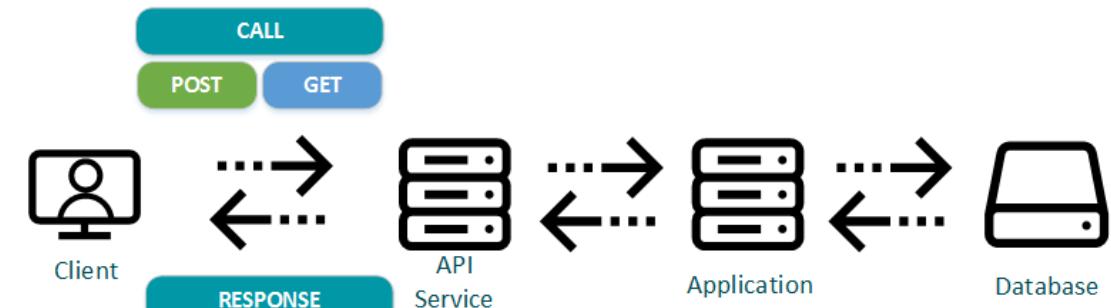
Mengambil Data Media Sosial

1. Mengambil di database internal
2. Melalui API
3. Scrapping
4. Screaming/Praying (Optional)

Twitter API

The screenshot shows the Twitter Developer Apps page. At the top, there's a purple navigation bar with links for Developer, Use cases, Products, Docs, More, Labs, Apps, dhitology, and a Create an app button. Below the navigation, a section titled 'Apps' lists two entries: 'Riset Kebijakan' and 'Thesis Dito'. Each entry includes a small Twitter icon, the app name, a blacked-out thumbnail, and a 'Details' button. At the bottom of the page, there are links for 'Developer policy and terms' and 'Follow @twitterdev', along with a 'Subscribe to developer news' button.

Membuat Twitter API: <http://apps.twitter.com/>



Sumber: https://help.eset.com/ema/2/api/en-US/how_api_works.html

rtweet

build passing CRAN 0.7.0 codecov 64% DOI 10.5281/zenodo.2528481 R Peer Reviewed
 repo status Active downloads 13K/month downloads 180K lifecycle maturing JOSS 10.21105/joss.01829

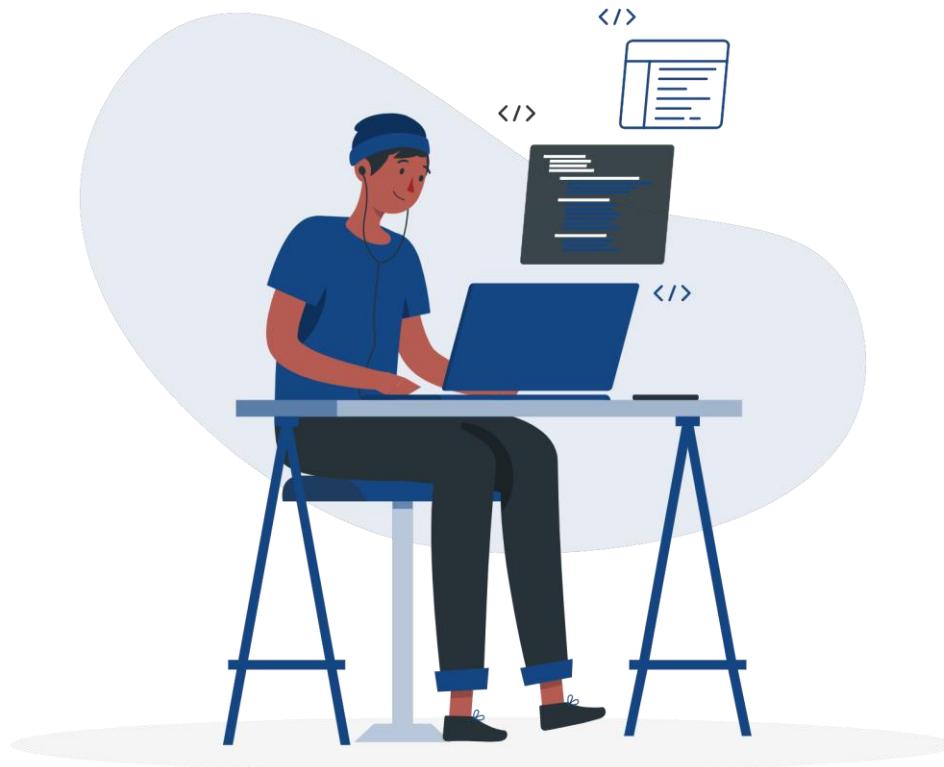
R client for accessing Twitter's REST and stream APIs. Check out the [rtweet package documentation website](#).

Task	rtweet	twitteR	streamR	RTwitterA
Available on CRAN	✓	✓	✓	✗
Updated since 2016	✓	✗	✓	✗
Non-'developer' access	✓	✗	✗	✗
Extended tweets (280 chars)	✓	✗	✓	✗
Parses JSON data	✓	✓	✓	✗
Converts to data frames	✓	✓	✓	✗
Automated pagination	✓	✗	✗	✗
Search tweets	✓	✓	✗	?
Search users	✓	✗	✗	?
Stream sample	✓	✗	✓	✗
Stream keywords	✓	✗	✓	✗
Stream users	✓	✗	✓	✗
Get friends	✓	✓	✗	✓
Get timelines	✓	✓	✗	?
Get mentions	✓	✓	✗	?
Get favorites	✓	✓	✗	?



Rtweeet

```
[1] "user_id"                               "status_id"
[4] "screen_name"                           "text"
[7] "display_text_width"                   "reply_to_status_id"
[10] "reply_to_screen_name"                "is_quote"
[13] "favorite_count"                     "retweet_count"
[16] "reply_count"                         "hashtags"
[19] "urls_url"                            "urls_t.co"
[22] "media_url"                           "media_t.co"
[25] "media_type"                          "ext_media_url"
[28] "ext_media_expanded_url"             "ext_media_type"
[31] "mentions_screen_name"               "lang"
[34] "quoted_text"                         "quoted_created_at"
[37] "quoted_favorite_count"              "quoted_retweet_count"
[40] "quoted_screen_name"                 "quoted_name"
[43] "quoted_friends_count"               "quoted_statuses_count"
[46] "quoted_description"                  "quoted_verified"
[49] "retweet_text"                        "retweet_created_at"
[52] "retweet_favorite_count"              "retweet_retweet_count"
[55] "retweet_screen_name"                 "retweet_name"
[58] "retweet_friends_count"              "retweet_statuses_count"
[61] "retweet_description"                 "retweet_verified"
[64] "place_name"                          "place_full_name"
[67] "country"                            "country_code"
[70] "coords_coords"                      "bbox_coords"
[73] "name"                                "location"
[76] "url"                                  "protected"
[79] "friends_count"                       "listed_count"
[82] "favourites_count"                    "account_created_at"
[85] "profile_url"                         "profile_expanded_url"
[88] "profile_banner_url"                  "profile_background_url"
                                                "profile_image_url"
```

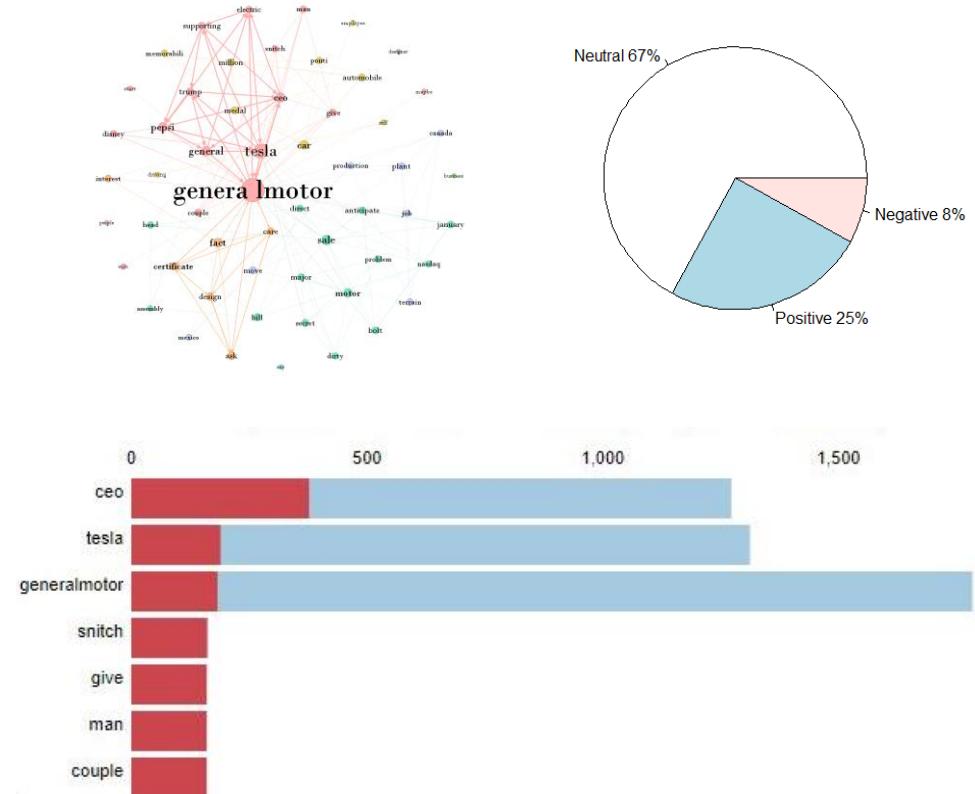


Yuk, Praktek

Biar lebih ngerti

Text Mining

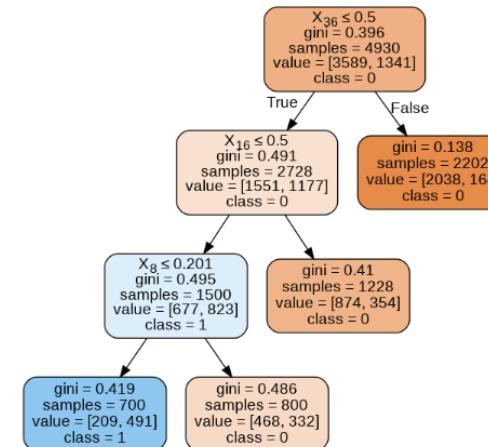
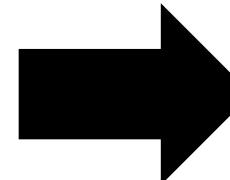
- **Text Mining** merupakan proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipan teks, dll. (*Wikipedia*)
 - **Natural Language Processing (NLP)** merupakan cabang ilmu komputer dan linguistik yang mengkaji interaksi antara komputer dengan bahasa (alami) manusia. NLP merupakan salah satu komponen Text Mining yang melakukan analisis linguistik untuk membantu mesin agar "membaca" teks.



Text Mining dan Data Mining

Data Mining

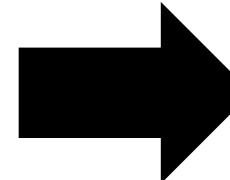
customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
7590-VHVEG	Female	No	Yes	No	1	No	No phone service	DSL
5575-GNVDDE	Male	No	No	No	34	Yes	No	DSL
3668-QPYBK	Male	No	No	No	2	Yes	No	DSL
7795-CFOCW	Male	No	No	No	45	No	No phone service	DSL
9237-HQITU	Female	No	No	No	2	Yes	No	Fiber optic
9305-CDSKC	Female	No	No	No	8	Yes	Yes	Fiber optic
1452-KIOVK	Male	No	No	Yes	22	Yes	Yes	Fiber optic
6713-OKOMC	Female	No	No	No	10	No	No phone service	DSL
7892-POOKP	Female	No	Yes	No	28	Yes	Yes	Fiber optic
6388-TABGU	Male	No	No	Yes	62	Yes	No	DSL
9763-GRSKD	Male	No	Yes	Yes	13	Yes	No	DSL
7469-LKBCI	Male	No	No	No	16	Yes	No	No
8091-TTVAX	Male	No	Yes	No	58	Yes	Yes	Fiber optic
0280-XJGEX	Male	No	No	No	49	Yes	Yes	Fiber optic
5129-JLPIS	Male	No	No	No	25	Yes	No	Fiber optic



Text Mining

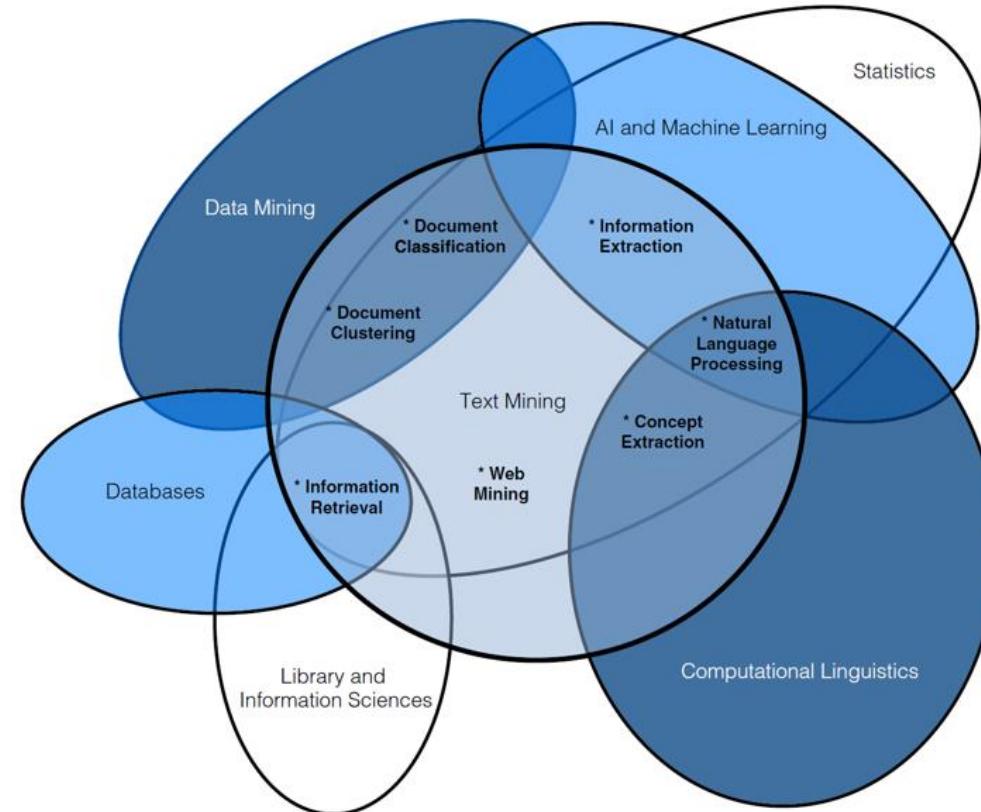
Sebelum ricuh, perwakilan mahasiswa sudah diterima pimpinan DPRD untuk menyampaikan aspirasinya. Aparat keamanan kemudian mengimbau massa agar tak mendesak masuk gedung dewan. Namun imbauan itu tak diindahkan mahasiswa.

Bersamaan dengan lemparan botol, batu, sepatu dan barang-barang lainnya, mahasiswa yang berada depan pagar merangsek masuk. Pagar yang tadinya menyekat mahasiswa dengan aparat keamanan akhirnya jebol.



Important Keyword:
 Ricuh
 Mahasiswa
 DPR
 Aparat
 Desak
 Masuk

The Big Picture



Sumber: The Seven Practice Areas of Text Analytics

Aplikasi Text Mining

Manufacturers

- Identify root causes of product issues quicker
- Identify trends in market segments
- Understand competitors' products

Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy

Financial Institutions

- Use contact center transcriptions understand customers
- Identify money laundering or other fraudulent situations

Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media

Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications

Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect disease outbreaks earlier
- Identify patterns in patient claims data

Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments

Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials

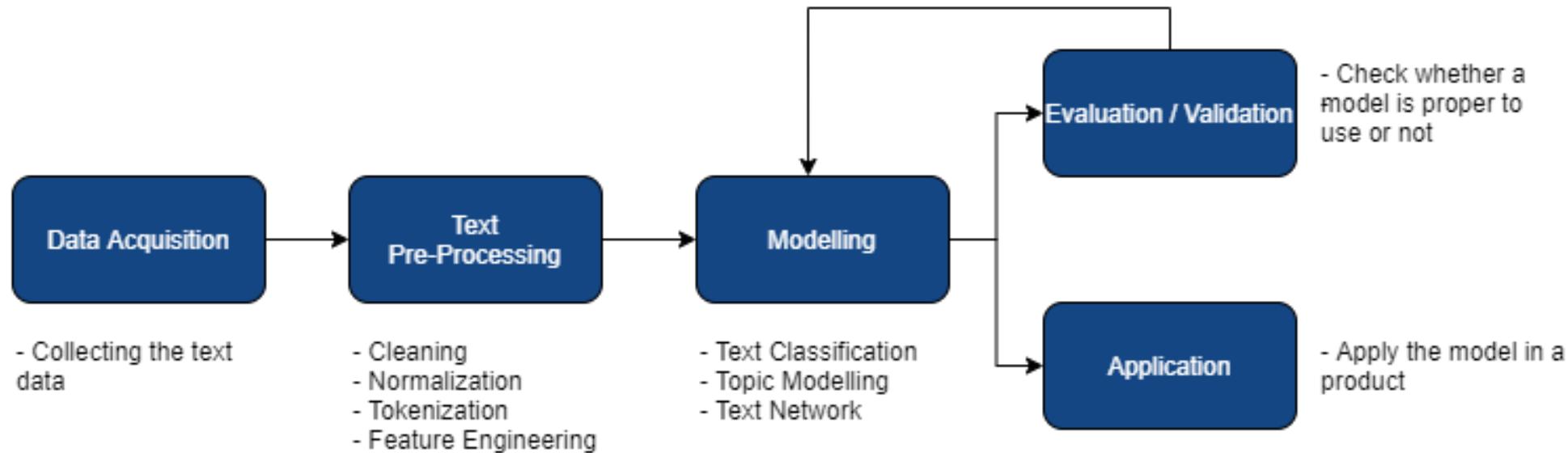
Insurance

- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos

Sumber: <https://www.zencos.com/blog/text-mining-examples-advanced-analytics/>

Proses Text Mining



Wordcloud

Wordcloud merupakan salah satu teknik visualisasi kata untuk melihat frekuensi masing-masing kata pada keseluruhan dokumen.

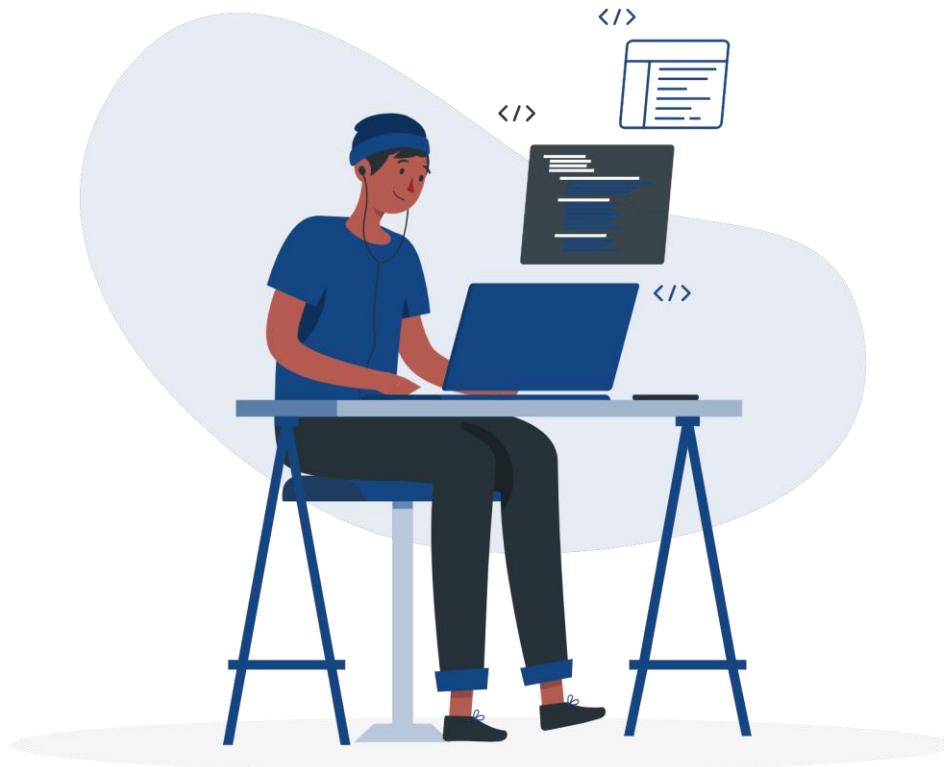
	word	freq
bahasa	bahasa	18.790670
indonesia	indonesia	18.136450
trending	trending	14.633601
pesona	pesona	11.183825
orang	orang	10.488977
raya	raya	10.100387
aceh	aceh	10.010657
sabar	sabar	8.961507
kisah	kisah	8.434360
negara	negara	8.166458



Text Network Analytics

- **Text Network Analysis** memetakan hubungan antar kata yang dapat memberikan gambaran besar terkait isi dari sebuah dokumen.





Yuk, Praktek

Biar lebih ngerti

Text Classification

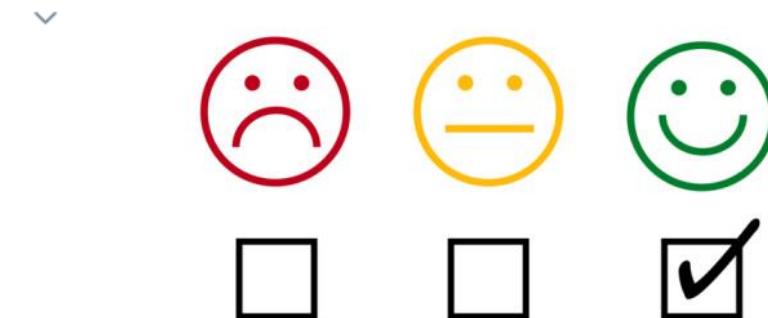
Text Classification atau klasifikasi text merupakan proses mengkategorikan dokumen atau kalimat kedalam sebuah label kelas tertentu. Contoh penerapan Text classification yang umum digunakan adalah Sentiment Analysis.



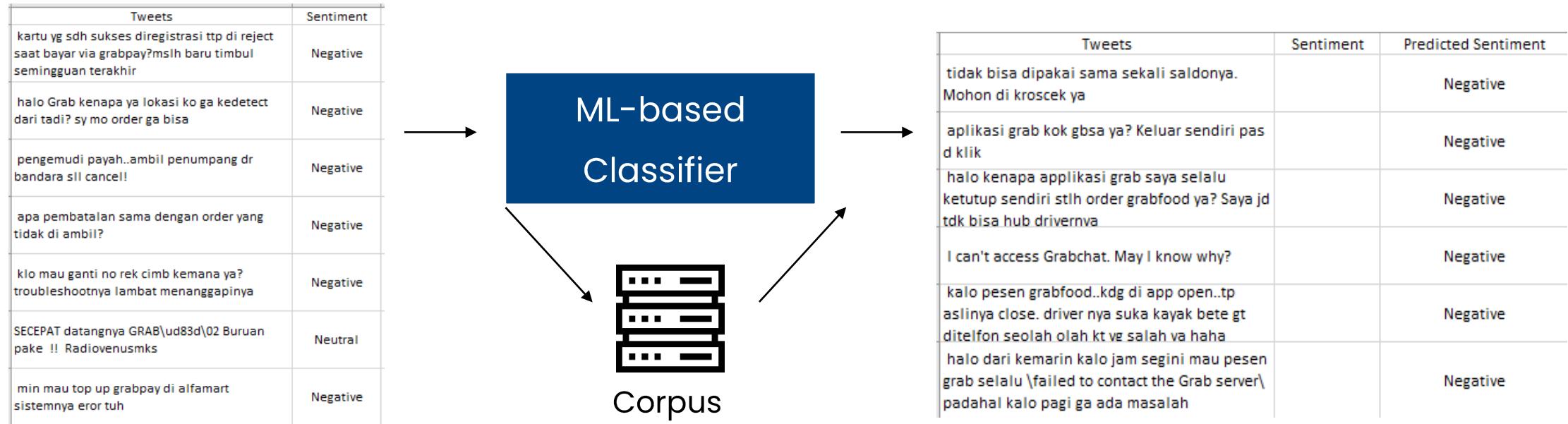
Ini dari tadi naik grab, abangnya wangi2 dan ramah ramah,

Bagus nih pelayanan grab di solo 🌟🌟 @GrabID

1.28 PM · 17 Okt 2018 · Twitter for Android



Sentiment Analysis



Problem 1: Uncleaned Text

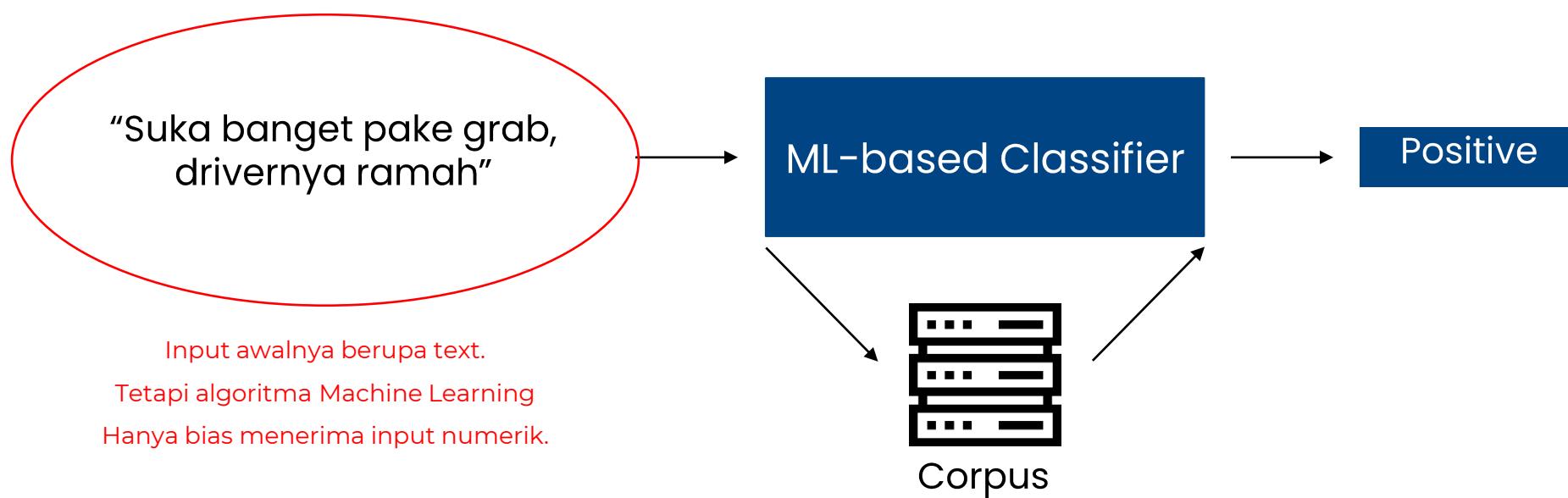


siam 18 @mimiamiamia95 · 3m

O gitu caranya banya followers bikin GA ipon, pas udah banyak yang follow GAnyya gajadi diumumin deh atau bilangnya udah ada yang menang hhhhhh
Bisa saja **klean** ini ngibulnya 😊😊

Tipe Pre-Processing	Hasil
Tokenization	O gitu caranya banya followers bikin GA ipon, pas sudah banyak yang follow GAnyya gajadi diumumin deh atau bilangnya sudah ada yang menang hhhhhh Bisa saja klean ini ngibulnya
Slang word	O gitu caranya banyak followers bikin GA iphone, pas sudah banyak yang follow GAnyya tidak jadi diumumin deh atau bilangnya sudah ada yang menang hhhhhh Bisa saja kalian ini ngibulnya
Stemming	O gitu cara banyak followers bikin GA iphone, pas sudah banyak yang follow GA tidak jadi umum deh atau bilang sudah ada yang menang hhhhhh Bisa saja kalian ini ngibul
Lemmatization	O gitu cara banyak followers buat GA iphone, pas sudah banyak yang follow GA tidak jadi umum deh atau sebut sudah ada yang menang hhhhhh Bisa saja kalian ini tipu
Stop word	cara banyak followers GA iphone, banyak follow GA tidak umum sebut menang kalian tipu.

Problem 2: Numerical Input



Text Vectorization

D = aku suka menggunakan grab karena murah

Tokenize

(“aku”, “suka ”, “menggunakan”, “grab” , “karena”, “murah”)



D = X1, X2, X3, ... Xn

$W = \text{Word (Text)}$

$X = \text{Some numeric encoding of Word}$

Text Vectorization

Teknik Popular:

- One-Hot Encoding
- Frequency-Based Encoding: (TF-IDF, Co-occurrence)
- Prediction-Based Encoding: (Lda2Vec)

One Hot Encoding

Tweet
Direkomendasikan banget nih
Drivernya bau
Ramah banget
Gak bakal naik lagi

Semua Kata
Direkomendasikan
Banget
Nih
Drivernya
Bau
Ramah
Gak
Bakal
Naik
Lagi

One Hot Encoding

Semua Kata	Direkomendasikan banget nih	Drivernya bau	Ramah banget
direkomendasikan	1	0	0
banget	1	0	1
nih	1	0	0
drivernya	0	1	0
bau	0	1	0
ramah	0	0	1
gak	0	0	0
bakal	0	0	0
naik	0	0	0
lagi	0	0	0

Frequency Based Encoding

Teknik Popular

- **Count:** Melihat seberapa sering sebuah kata muncul di sebuah dokumen.
- **TF-IDF:** Melihat seberapa sering sebuah kata muncul di sebuah dokumen dan juga di keseluruhan dokumen.
- **Co-occurrence:** Kata yang sering mucncl Bersama memiliki konteks yang sama.

TF-IDF

TF-IDF: Melihat seberapa sering sebuah kata muncul di sebuah dokumen dan juga di keseluruhan corpus.

$$x_i = tf(w_i) \times idf(w_i)$$

tf = (Number of repetitions of word in a document) / (# of words in a document)

idf = Log[(# Number of documents) / (Number of documents containing the word)] and

TF-IDF

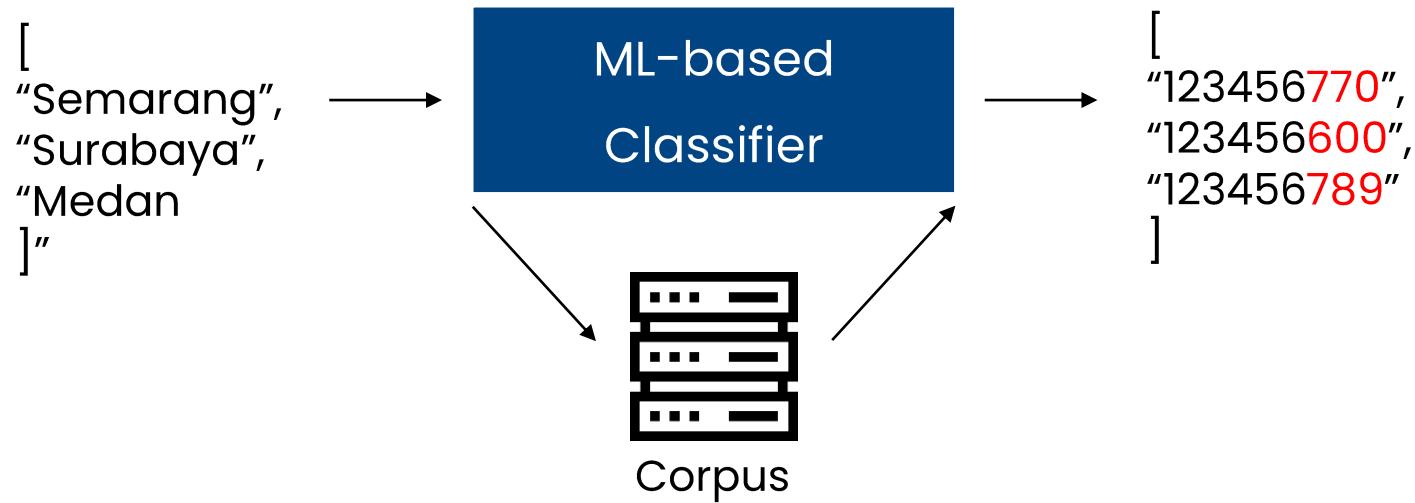


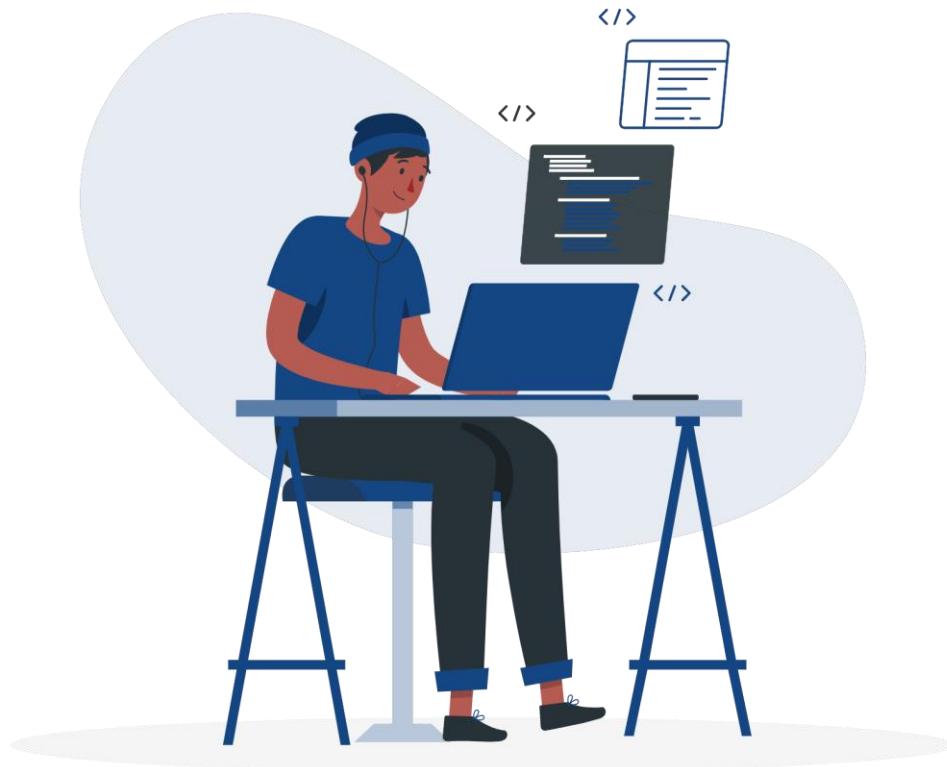
Semakin sering disebuah dokumen
Kemungkinan kata yang penting



Semakin sering muncul disemua dokumen
Kemungkinan kata yang biasa digunakan

Prediction Based Encoding





Yuk, Praktek

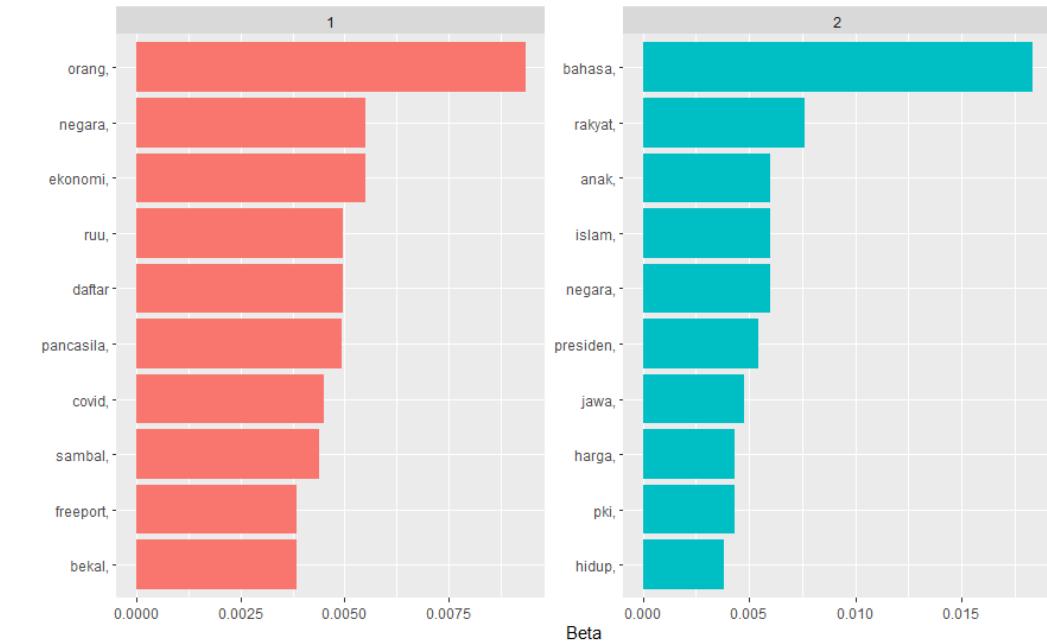
Biar lebih ngerti

Topic Modelling

Topic model merupakan salah satu metode statistic untuk mengekstraksi topik dari sebuah/kumpulan dokumen.

Teknik Popular:

- Latent Semantic Analysis (LSA)
- Probabilistic Latent Semantic Analysis (PLSA)
- **Latent Dirichlet Allocation (LDA)**
- Lda2vec



Latent Dirichlet Allocation

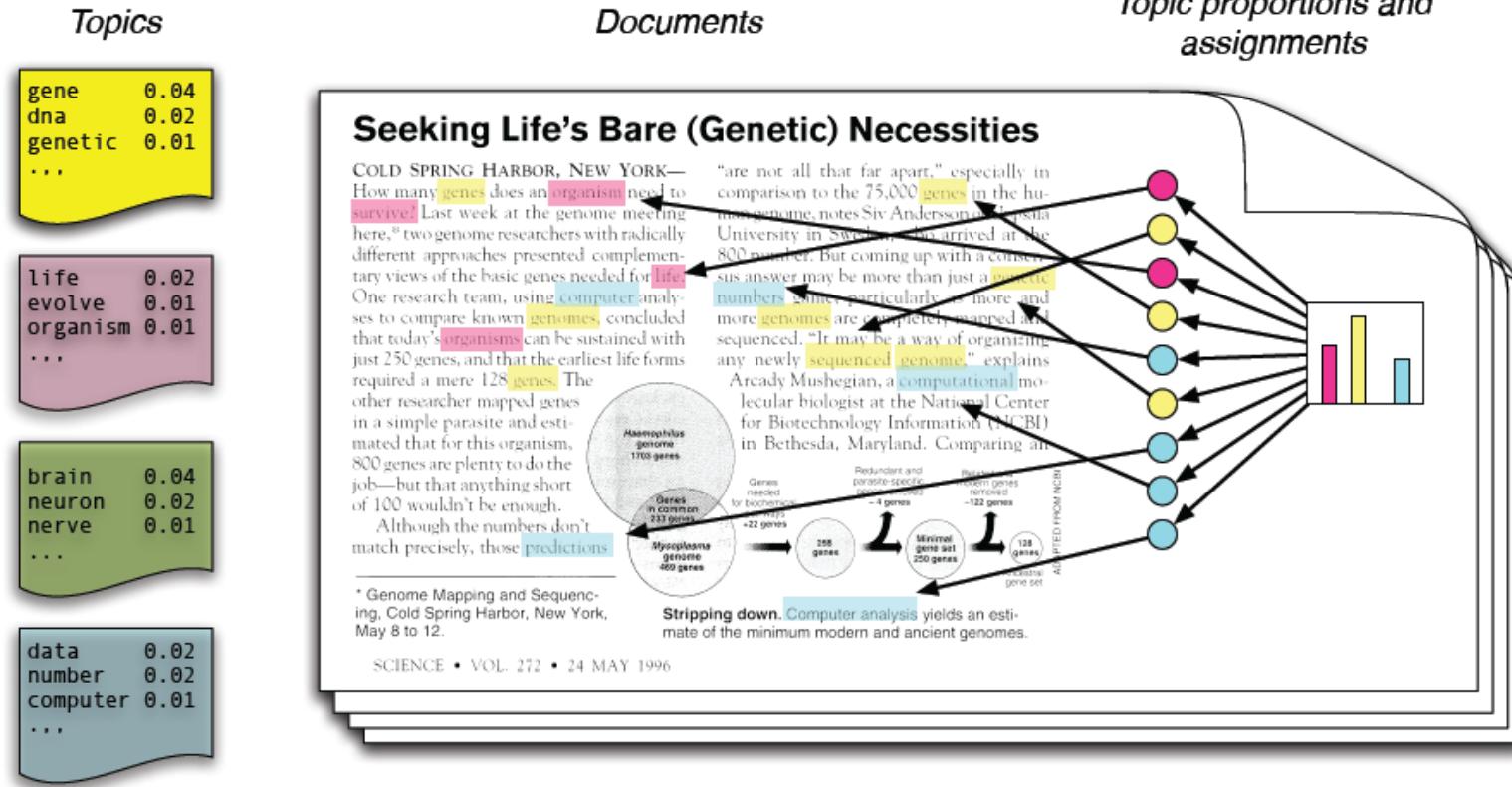
Asumsi dalam LDA:

- Setiap dokumen merupakan kumpulan Topik
- Setiap topik merupakan kumpulan kata

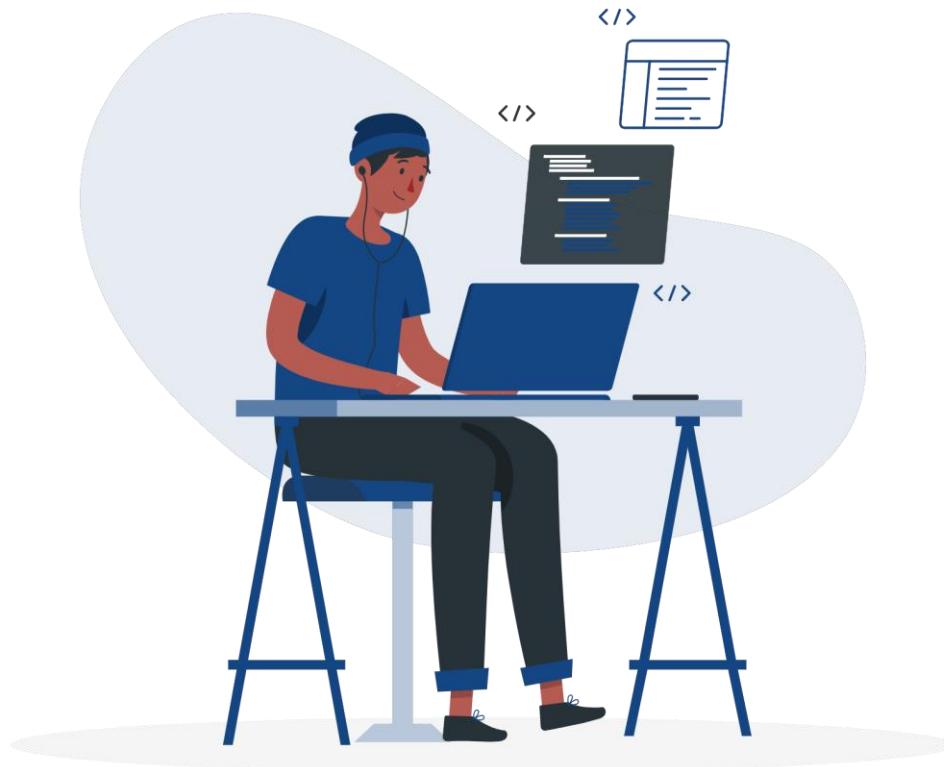
The news feed consists of five horizontal news cards. Each card includes a small thumbnail image, the word 'POLITICS' or 'ENTERTAINMENT' in red, the time '1 week ago' in grey, and a short black text summary.

- POLITICS** 1 week ago
House passes Perppu on CO' response amid concerns of embezzlement
- ENTERTAINMENT** 1 day ago
Writer says 'Extracurricular' digs into society's wounds
- POLITICS** 1 week ago
PAN's internal rifts linger after congress violence, resignation of patron's son
- ENTERTAINMENT** 1 day ago
Rain is back: Singer's 2017 'Gang' video goes viral
- POLITICS** 1 week ago
PPATK deputy chief Dian Ediana Rae sworn in as new chairman
- ENTERTAINMENT** 1 day ago
Following protests, YG announces plans for BLACKPINK's first full-length album
- ENTERTAINMENT** 1 day ago
Ambyar Tak Jogesti concert to feature hologram of Didi Kempot
- POLITICS** 1 week ago
Jokowi issues Perppu to postpone 2020 regional elections amid outbreak

Latent Dirichlet Allocation



Sumber: Probabilistic Topic Models By David M. Blei

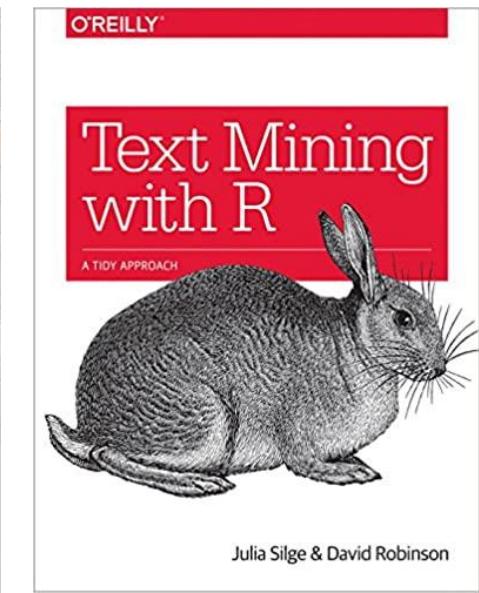
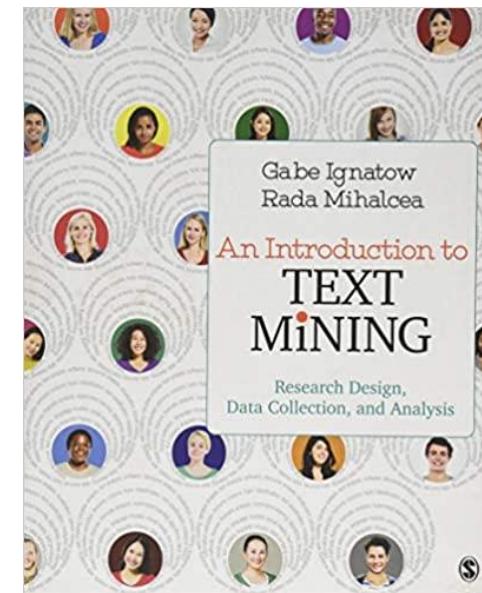


Yuk, Praktek

Biar lebih ngerti

Rekomendasi Buku

- An Introduction to Text Mining: Research Design, Data Collection, and Analysis 1st Edition
- Text Mining with R: A Tidy Approach 1st Edition



On Going Project

Input

Masukkan Keyword atau Hashtag yang ingin dianalisa:

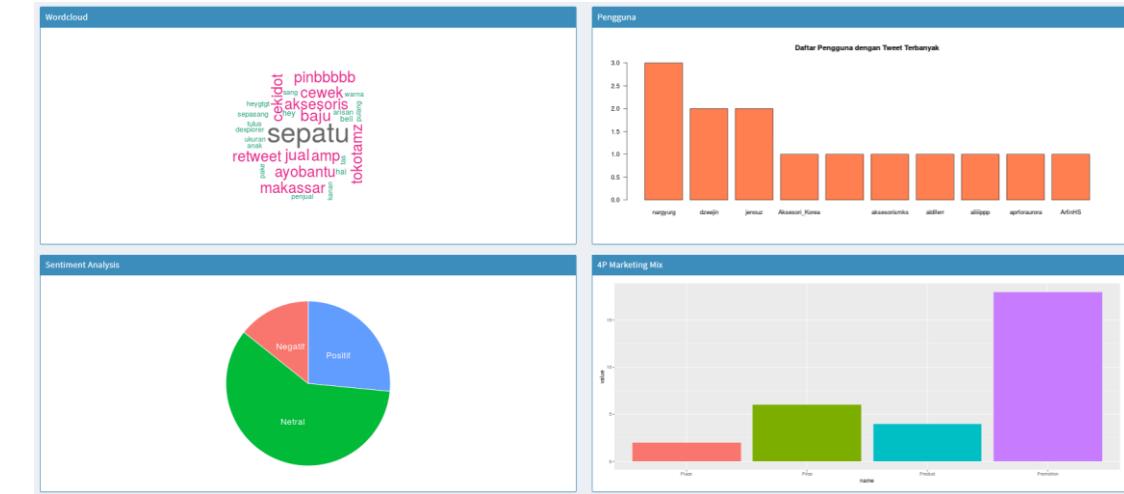
Jumlah Tweet:

10 100 300

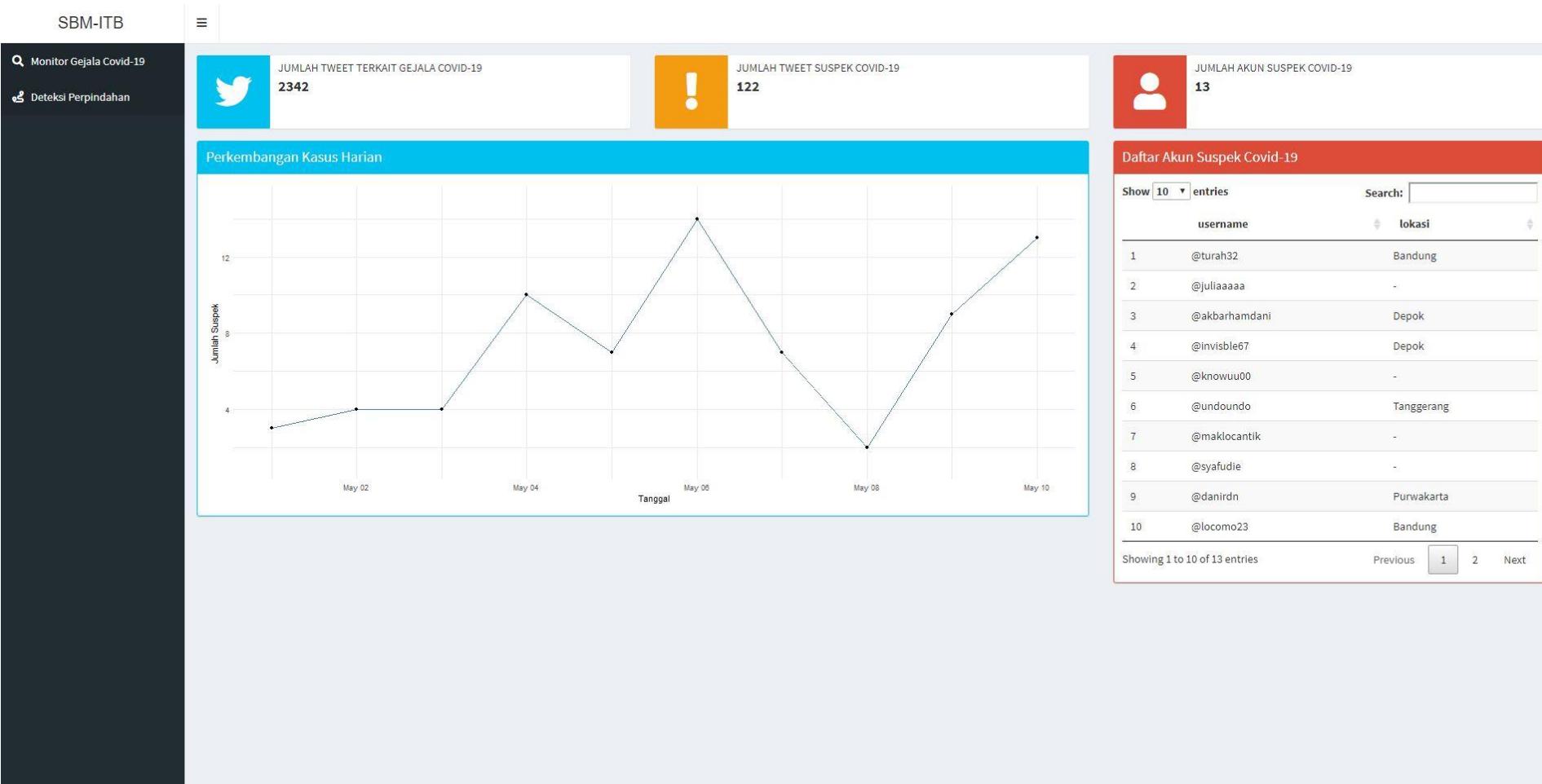
10 40 70 100 130 160 190 220 250 280 300

Termasuk Retweet

Analisa



On Going Project



Sekian, Terima Kasih

Jangan Lupa Bahagia ☺