# 3

# Collecting Social Media Data

Social Media Analytic Workshop

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# M. Aprian**dito** Arya Saputra



- Freelance Data Evangelist

- Former Data Scientist at Paperlust.Co

- Public Relations at Asosiasi Ilmuwan Data Indonesia

- Bachelor of Management, Telkom University

- Master of Business Administration Student, Institut Teknologi Bandung

**Phone**      **: +628112770063**
**Email**       **: m.apriandito@gmail.com**

# Learning Outline

- **Introduction**
- Collecting Data from Twitter
- Constructing Online Interaction Graph
- Collecting Data from Social Forum

# Collecting Data from Social Media

- First, the most direct way is to download the data from databases on the web servers.

- The second method is collecting data through application programming interfaces (APIs).

- The third method is web scraping. This method is particularly useful for those web-sites that do not provide APIs.

# Ethical issues

- Harvesting social media data may lead to an unwanted and unforeseen exposure of private data for participants .Large quantities of personal information on social media.

- Informed consent

- Participant anonymity

# Learning Outline

- Handshake with Social Media Data
- **Collecting Data from Twitter**
- Constructing Online Interaction Graph
- Collecting Data from Social Forum

# Practice: Collecting Twitter Data Using R



Practice Link: https://github.com/sma-workshop/r

# Practice: Collecting Twitter Data Using Python

Practice Link: https://github.com/sma-workshop/python

# Learning Outline

- Handshake with Social Media Data

- Collecting Data from Twitter

- **Constructing Online Interaction Graph**

- Collecting Data from Social Forum

# Build the Dataset

| screen_name<br>=<br>Source | reply_to_screen<br>=<br>Target |
|---|---|
| Dian | Rizqy |
| Dito | Rizqy |
| Dito | Dian |
|  |  |
| … | … |

# Practice: Constructing Online Interaction Graph

# Learning Outline

- Handshake with Social Media Data

- Collecting Data from Twitter

- Constructing Online Interaction Graph

- **Collecting Data from Social Forum**

# Practice: Collecting Review Data

# 4

# Text Mining

Social Media Analytic Workshop

Telkom University

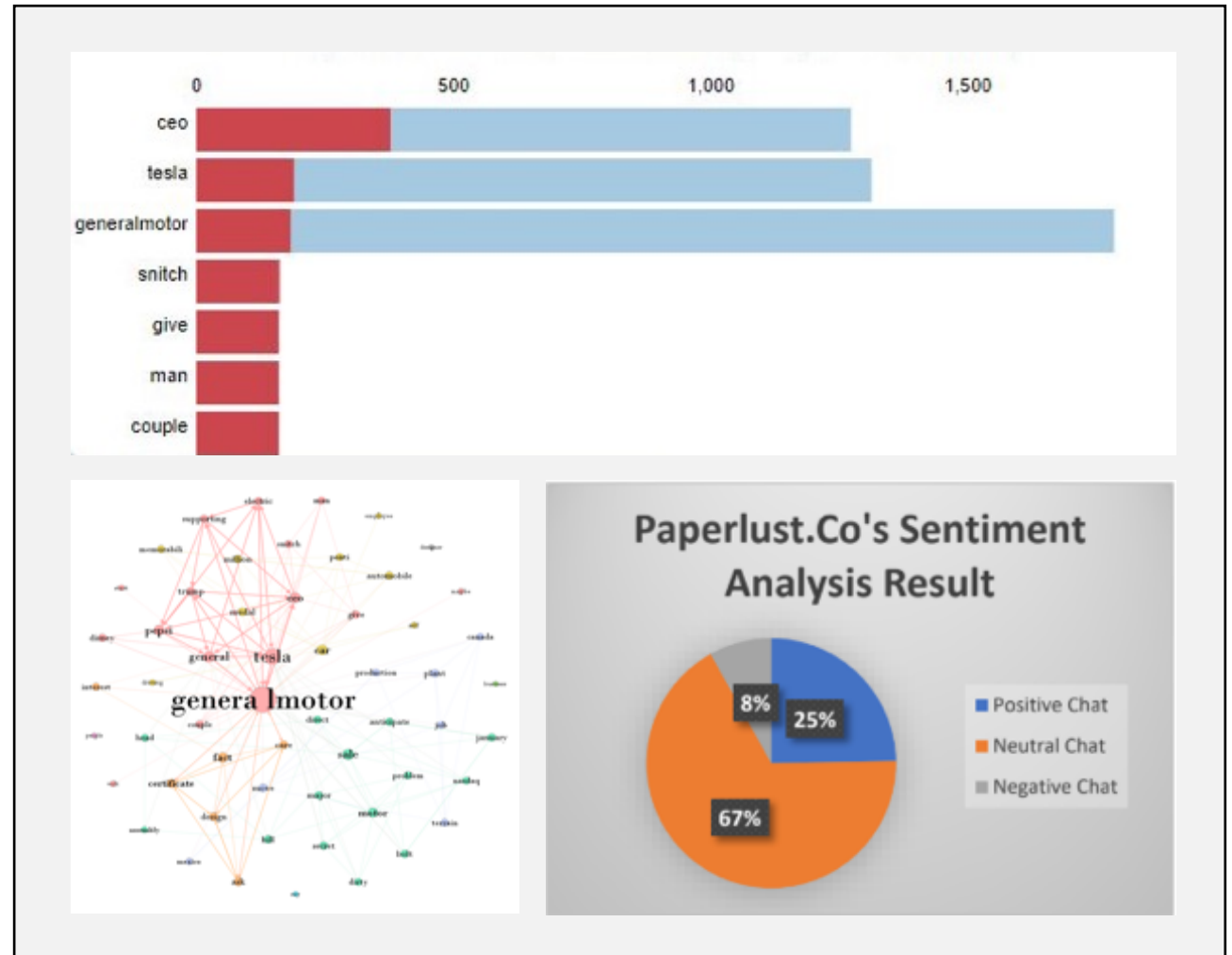DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Learning Outline

- **Introduction to Text Mining**

- Text Preprocessing

- Text Classification: Sentiment Analysis

- Topic Modelling

- Text Network Analysis

# Text Mining

- According to Wikipedia, "Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text."

- Text mining combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.
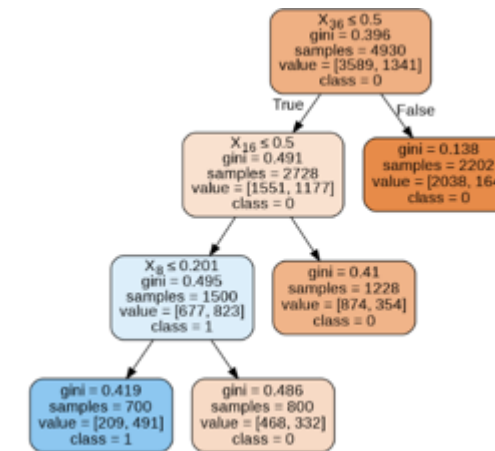


Paperlust.Co's Sentiment Analysis Result

# Data Mining and Text 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-GNVDE | 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 |

Output →



## 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.

Output →

```
Important Keyword:
Ricuh
Mahasiswa
DPR
Aparat
Desak
Masuk
```

# Learning Outline

- Introduction to Text Mining
- **Text Preprocessing**
- Text Classification: Sentiment Analysis
- Topic Modelling
- Text Network Analysis

# Everything was fine,
# until you meet Indonesian Language

**Kandiya R. Maulidita** ✔ @R_IndyJKT48 · 8h

Woi anak orang yang baik nya tulus amat! Habedeee nandskiiiiii mi amor😘
😘 Sampe setaun lbh ini gue kenal **lu**, gue gapernah ngeliat **lu** marah2 ato
sebel sm org lama bgtt... Sumpah gw gangerti **lu bisa** setulus itu baiknya😍
I LOP UU NANDSKI!! SUKSES SELALU YAW💖💖 #L16htenUpNandays

**siam** 🔞 @mimiamiamia95 · 3m

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

siam 🔞 @mimiamiamia95 · 3m
O gitu caranya banya followers bikin GA ipon, pas udah banyak yang pollow
GAnya gajadi diumumin deh atau bilangnya udah ada yang menang hhhhh
Bisa saja **klean** ini ngibulnya 😊😊

| Pre-Precessing Type | Result |
|---|---|
| Tokenization | O gitu caranya banya followers bikin GA ipon, pas sudah banyak yang pollow GAnya gajadi diumumin deh atau bilangnya sudah ada yang menang hhhhh Bisa saja klean ini ngibulnya |
| Slang word | O gitu caranya banyak followers bikin GA iphone, pas sudah banyak yang follow GAnya tidak jadi diumumin deh atau bilangnya sudah ada yang menang hhhhh 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 hhhhh 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 hhhhh Bisa saja kalian ini tipu |
| Stop word | cara banyak followers GA iphone, banyak follow GA tidak umum sebut menang kalian tipu. |

# Practice: Pre-Processing Text Data Using R



Practice Link: https://github.com/sma-workshop/r

# Learning Outline

- Introduction to Text Mining

- Text Preprocessing

- **Text Classification: Sentiment Analysis**

- Topic Modelling

- Text Network Analysis

# Text Classification

Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. Some of the most common examples and use cases for automatic text classification is **Sentiment Analysis.**



Ratna Astuti
@nhana_astuti

Ini dari tadi naik grab, abangnya wangi2 dan ramah ramah,
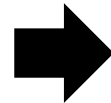Bagus nih pelayanan grab di solo 👌🙏 @GrabID
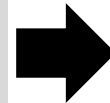
1.28 PM · 17 Okt 2018 · Twitter for Android

# Create Sentiment Analysis Model

It is using Machine Learning Principle

| Tweets | Sentiment |
|---|---|
| kartu yg sdh sukses diregistrasi ttp di reject saat bayar via grabpay?mslh baru timbul semingguan terakhir | Negative |
| halo Grab kenapa ya lokasi ko ga kedetect dari tadi? sy mo order ga bisa | Negative |
| pengemudi payah..ambil penumpang dr bandara sll cancel! | Negative |
| apa pembatalan sama dengan order yang tidak di ambil? | Negative |
| klo mau ganti no rek cimb kemana ya? troubleshootnya lambat menanggapinya | Negative |
| SECEPAT datangnya GRAB\ud83d\02 Buruan pake !! Radiovenusmks | Neutral |
| min mau top up grabpay di alfamart sistemnya eror tuh | Negative |

→ **Machine Learning Algorithm** →

| Tweets | Sentiment | Predicted Sentiment |
|---|---|---|
| tidak bisa dipakai sama sekali saldonya. Mohon di kroscek ya | | Negative |
| aplikasi grab kok gbsa ya? Keluar sendiri pas d klik | | Negative |
| halo kenapa applikasi grab saya selalu ketutup sendiri stlh order grabfood ya? Saya jd tdk bisa hub drivernva | | Negative |
| I can't access Grabchat. May I know why? | | Negative |
| kalo pesen grabfood..kdg di app open..tp aslinya close. driver nya suka kayak bete gt ditelfon seolah olah kt vg salah ya haha | | Negative |
| halo dari kemarin kalo jam segini mau pesen grab selalu \failed to contact the Grab server\ padahal kalo pagi ga ada masalah | | Negative |

# Practice: Sentiment Analysis Using R



Practice Link: https://github.com/sma-workshop/r

# Practice: Sentiment Analysis Using Python
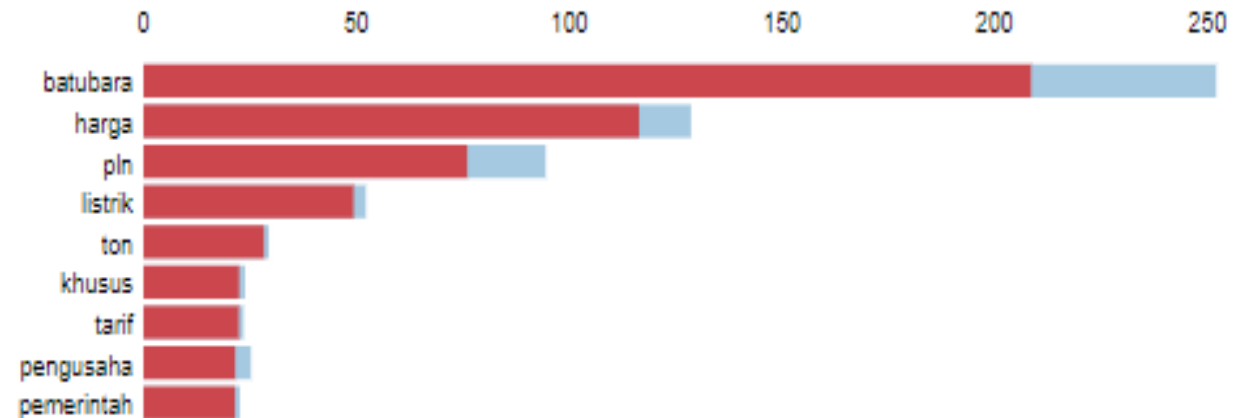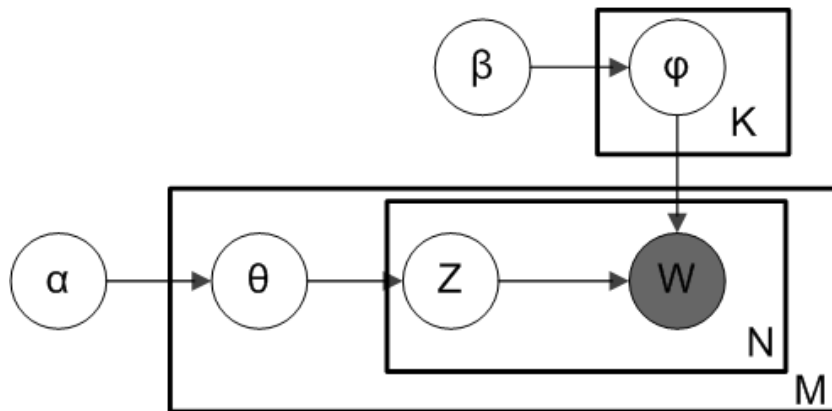


Practice Link: https://github.com/sma-workshop/python

# Learning Outline

- Introduction to Text Mining
- Text Preprocessing
- Text Classification: Sentiment Analysis
- **Topic Modelling**
- Text Network Analysis

# Topic Modelling

- Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents..

- LDA is easily the most popular (and typically most effective) topic modeling technique out there.

# Practice: LDA Topic Modelling Using Python



Practice Link: https://github.com/sma-workshop/python

# Practice: LDA Topic Modelling Using R



Practice Link: https://github.com/sma-workshop/r

# Learning Outline

- Introduction to Text Mining
- Text Preprocessing
- Text Classification: Sentiment Analysis
- Topic Modelling
- **Text Network Analysis**

# Text Network Analysis

- The analysis of text describing various kinds of "computer support solution" that allows the analyst to "extract networks of concepts" from text

- The assumption underlying the emergence text network is a relationship that is contained in every word so that it produces a pattern that can be analyzed

# Practice: Text Network Analysis

See You Again at
**Telkom University**