TELKOM
DIGITAL TALENT
INCUBATOR 2020

Modul 7: Text Mining

Organized by:
CAE

Module Overview

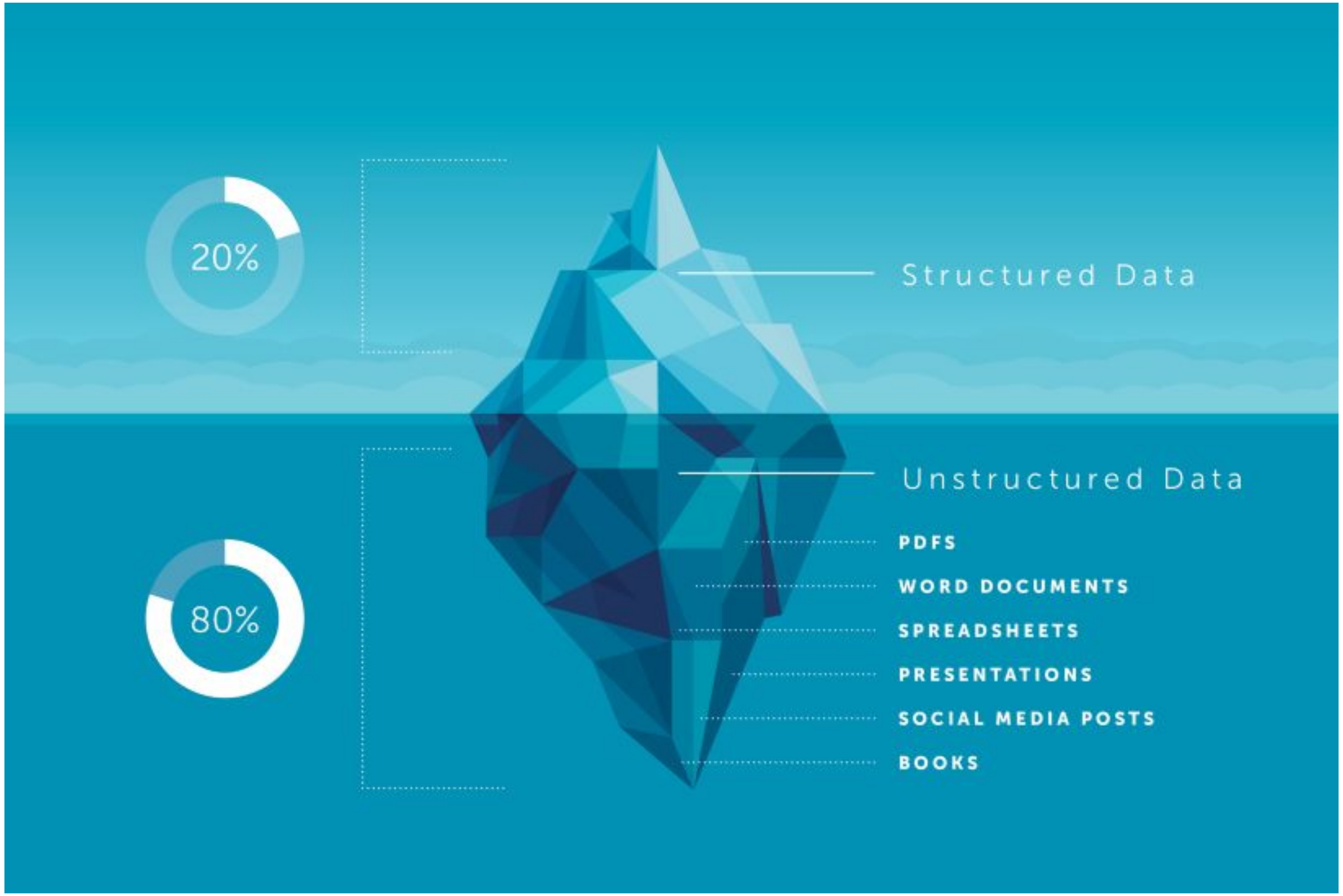
Topics

- What is Text Mining
- Text Mining Process
- Text Classification
- Topic Modelling

Activities

- Coding Practice
- Orange

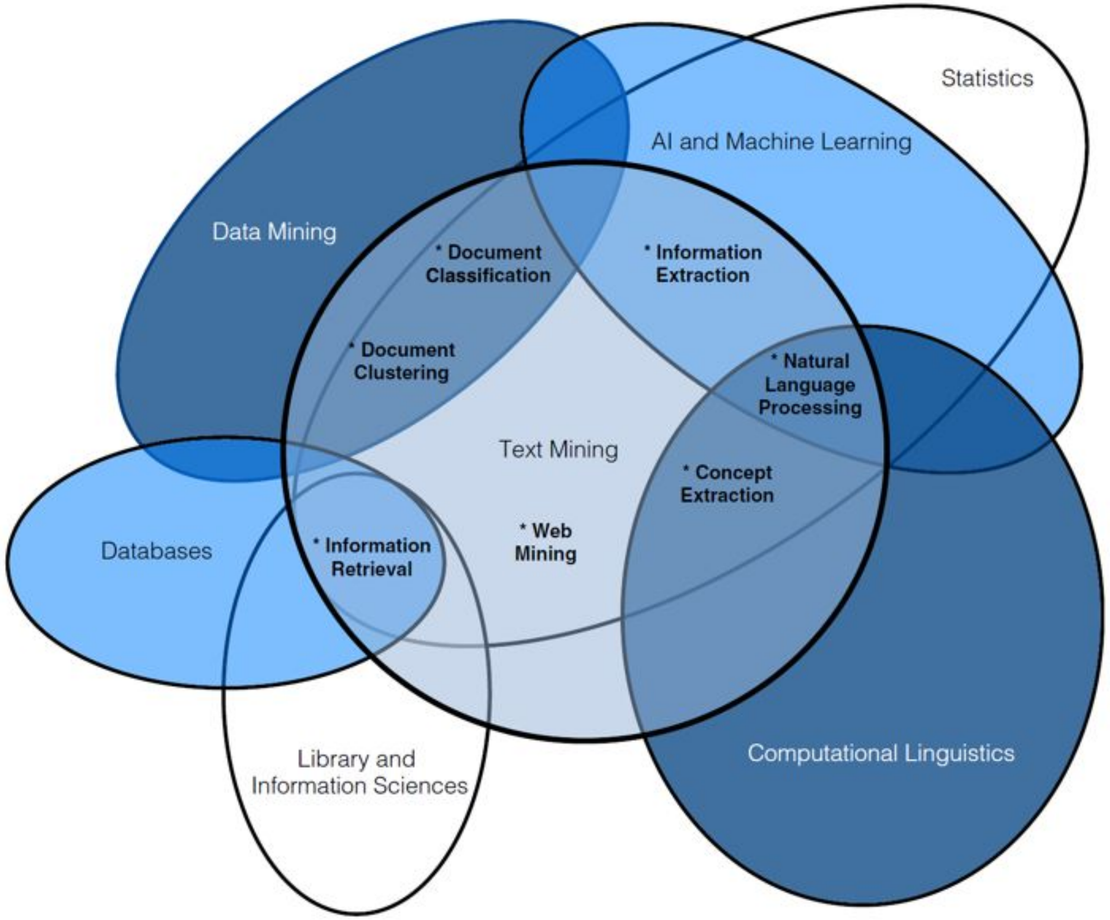
Data In The World



Text Mining

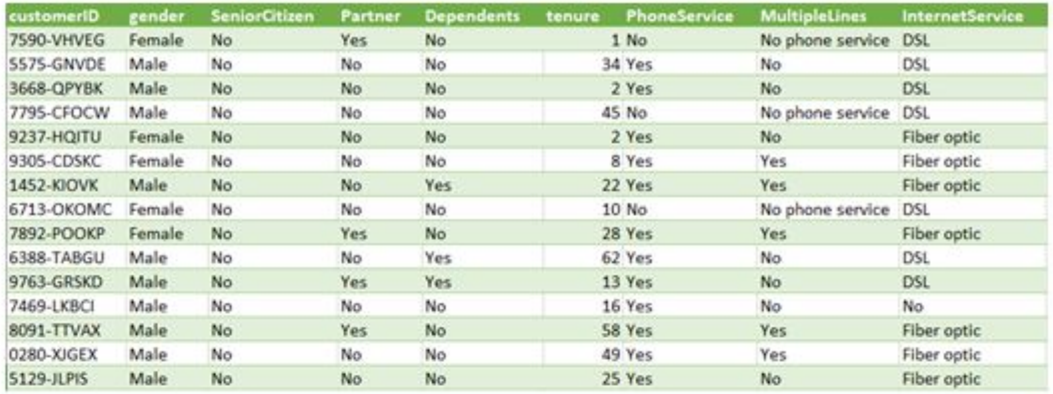
- Text mining is the process of extracting the implicit knowledge from textual data
- Because the implicit knowledge which is the output of text mining does not exist in the given storage, it should be distinguished from the information which is retrieved from the storage.

Text Mining

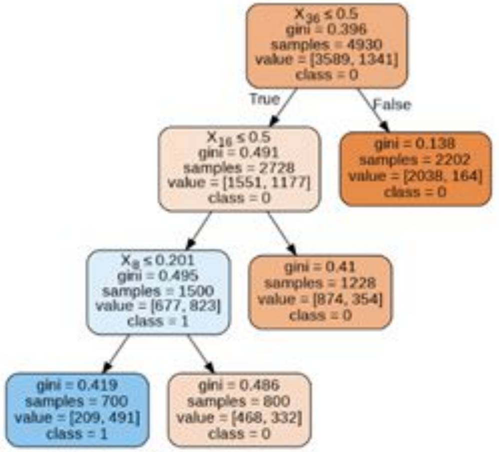

Statistics
AI and Machine Learning
Data Mining
* Document Classification
* Information Extraction
* Document Clustering
* Natural Language Processing
Text Mining
* Concept Extraction
Databases
* Information Retrieval
* Web Mining
Library and Information Sciences
Computational Linguistics

Text Mining and Data Mining

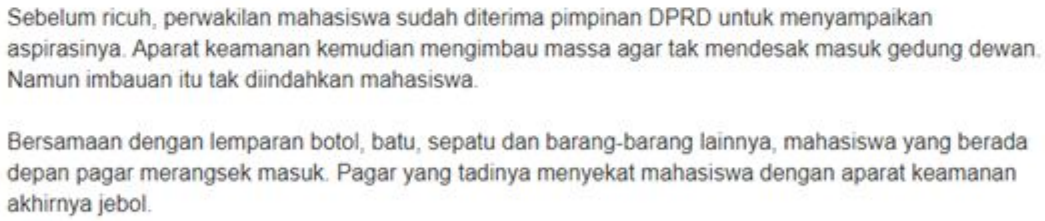
Data Mining



customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
7590-VHVEG	Female	No	Yes	No	1	No	No phone service	DSL
5575-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

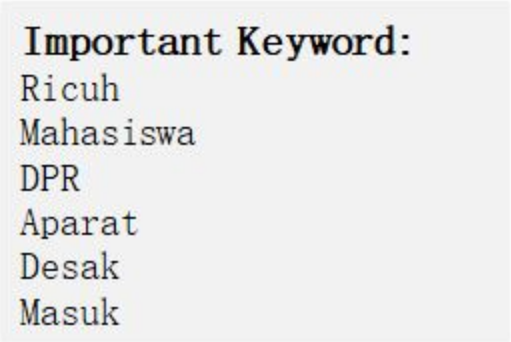


Text Mining



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.



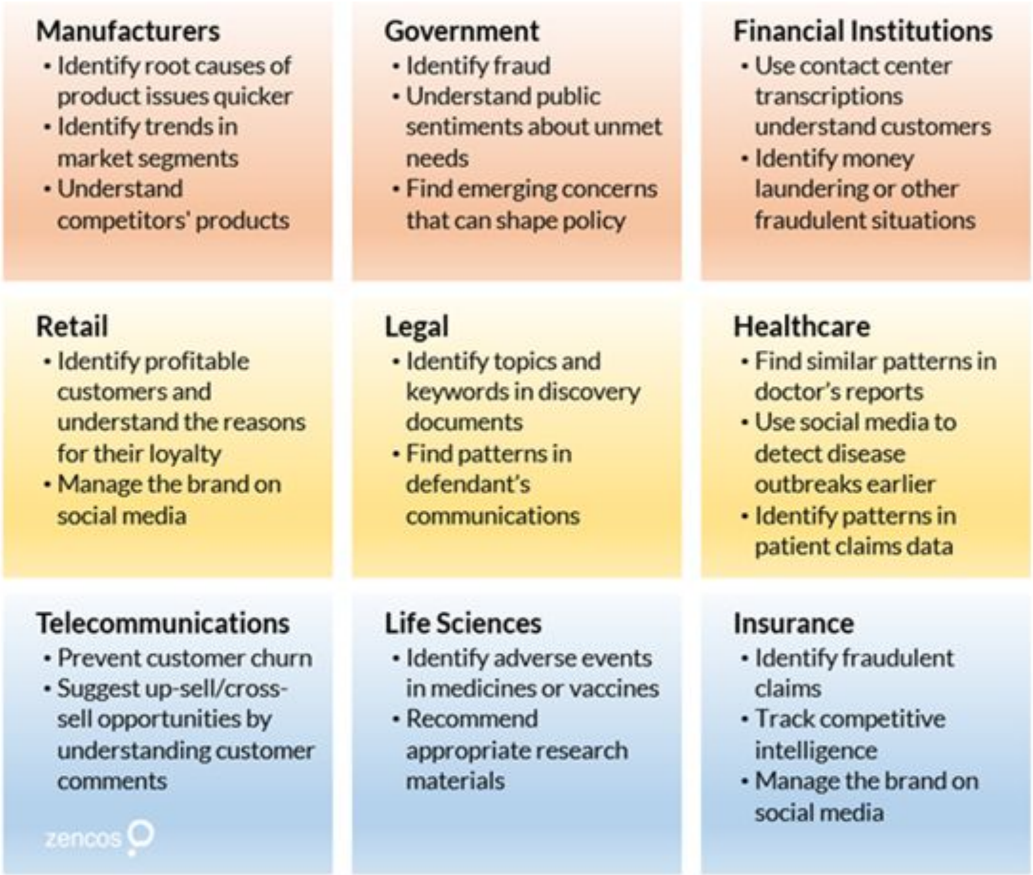
Important Keyword:
Ricuh
Mahasiswa
DPR
Aparat
Desak
Masuk

Text Mining Method

Retrieval	Obtaining information resources relevant to an information need from a collection of information resources
Extraction	Extracting structured information from unstructured and/or semi-structured machine-readable documents.
Summarization	process of automatically generating a compressed version of a specific text that holds valuable information for the end-user.
Categorization	Assigned Document to a predefined set of topics depending upon their content.
Cluster	identify intrinsic structures in textual information and organize them into relevant subgroups or ‘clusters’ for further analysis.

TELKOM DIGITAL TALENT INCUBATOR 2020

Text Mining Application



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

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

Organized by:

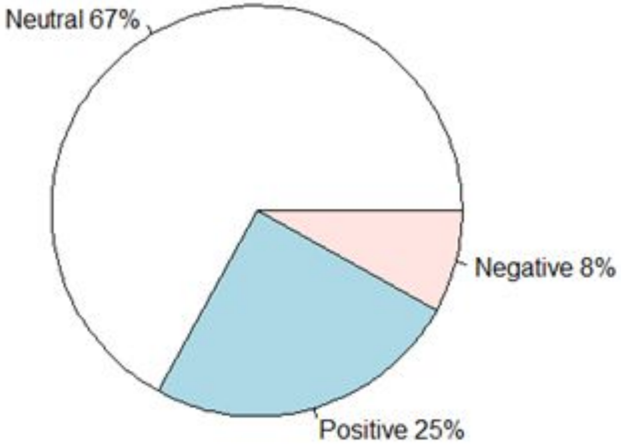
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

My Previous Project

- Customer Satisfaction Analysis in E-commerce

No.	Customer Chat	Sentiment Class
1.	I'm having a look now Thank you so much for your help You were amazing	Positive
2.	Will you be getting a sample pack of this design in?	Neutra
3.	This is not addressing again at all how bad of experience i have had the delays in my printing and approval the service horrific	Negative

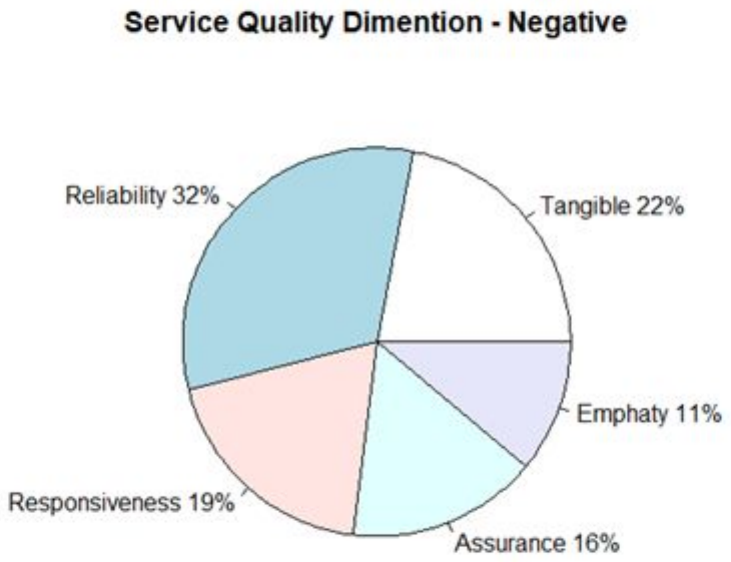


Organized by:
CAE

My Previous Project

- Customer Satisfaction Analysis in E-commerce

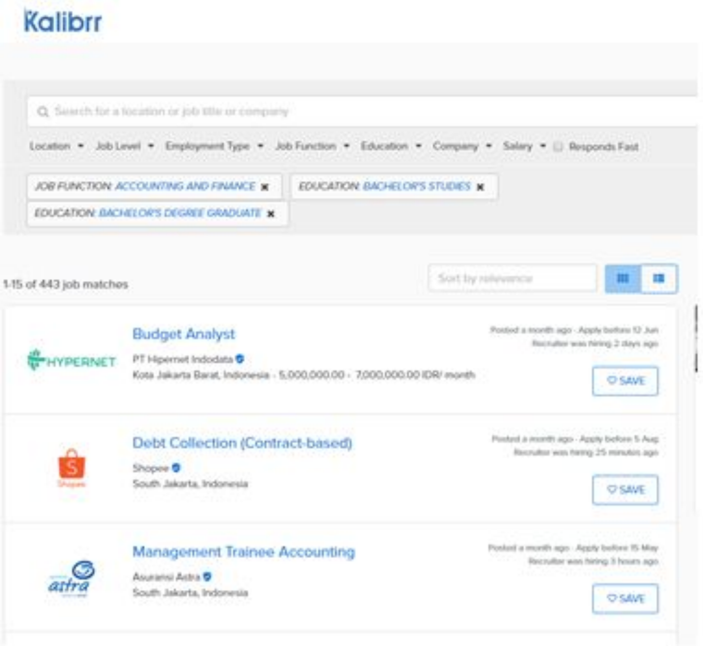
No.	Negative Customer Chat	Service Quality Dimension
1.	Flow on this web is confusing, I can't find what I'm looking for	Tangible
2.	I tried to contact Jamie to discuss this design problem, but he gave no response.	Responsiveness
3.	This is not addressing again at all how bad of experience I have had the delays in my printing and approval the service horrific	Reliability



TELKOM
DIGITAL TALENT
INCUBATOR 2020

My Previous Project

- Job Opportunities for Business School Graduates



Kalibrr

Search for a location or job title or company

Location ▾ Job Level ▾ Employment Type ▾ Job Function ▾ Education ▾ Company ▾ Salary ▾ □ Responds Fast

JOB FUNCTION: ACCOUNTING AND FINANCE ✖ EDUCATION: BACHELOR'S STUDIES ✖
EDUCATION: BACHELOR'S DEGREE GRADUATE ✖

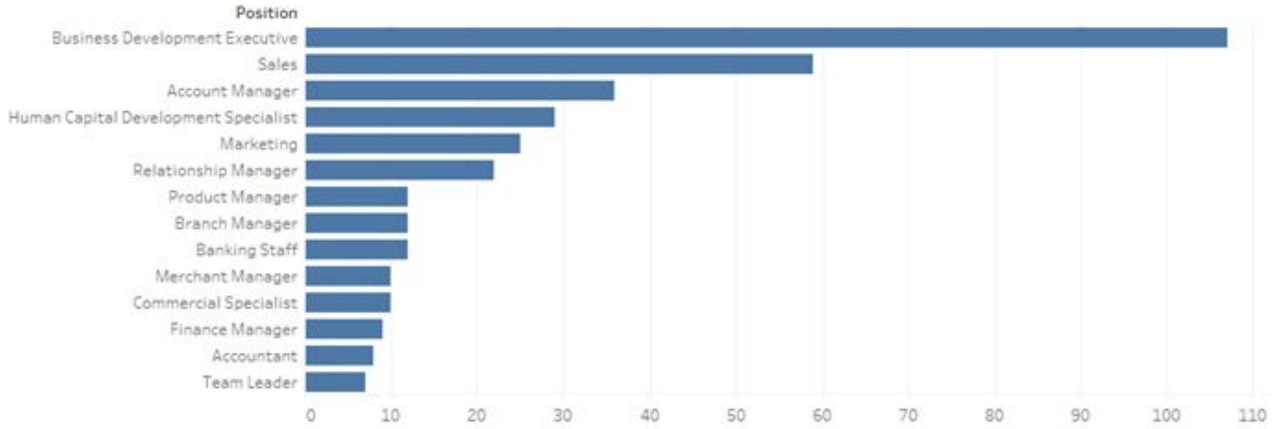
1-15 of 443 job matches

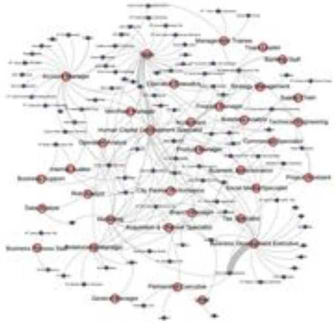
Sort by relevance

Budget Analyst
PT Hipernet Indodata
Kota Jakarta Barat, Indonesia · 5,000,000.00 - 7,000,000.00 IDR/ month
Posted a month ago · Apply before 12 Jun
Recruiter was hiring 2 days ago

Debt Collection (Contract-based)
Shopee
South Jakarta, Indonesia
Posted a month ago · Apply before 5 Aug
Recruiter was hiring 25 minutes ago

Management Trainee Accounting
Asuransi Astra
South Jakarta, Indonesia
Posted a month ago · Apply before 15 May
Recruiter was hiring 3 hours ago



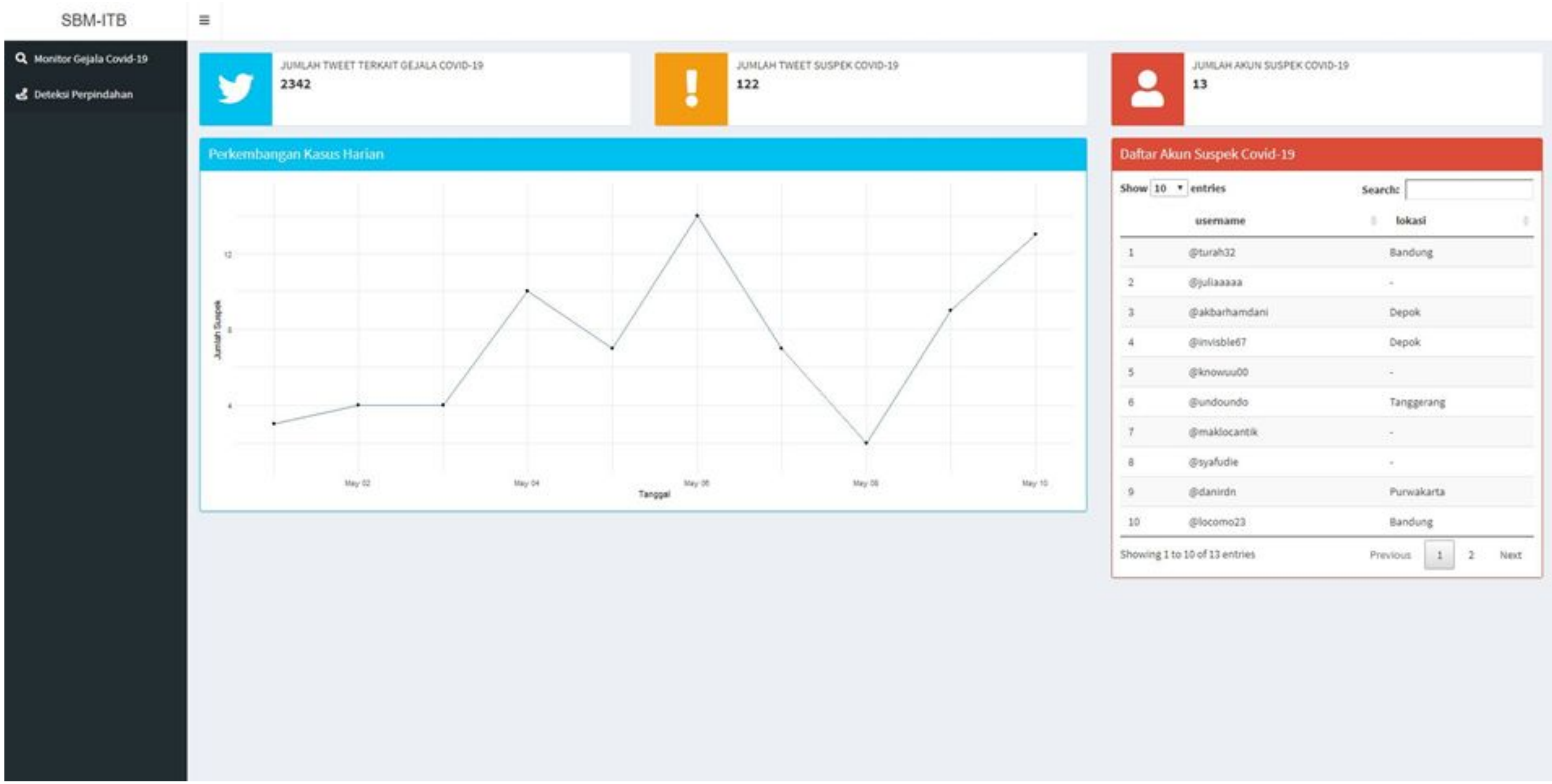


Rata-Rata Penawaran Gaji Terendah	Rata-Rata Penawaran Gaji Teringgi
Rp. 5,879,736	Rp. 8,394,930

Organized by:
CAE

My Previous Project

- Twitter COVID-19 Detection



SBM-ITB

Monitor Gejala Covid-19

Deteksi Perpindahan

JUMLAH TWEET TERKAIT GEJALA COVID-19
2342

JUMLAH TWEET SUSPEK COVID-19
122

JUMLAH AKUN SUSPEK COVID-19
13

Perkembangan Kasus Harian

Daftar Akun Suspek Covid-19

Show 10 entries

Search:

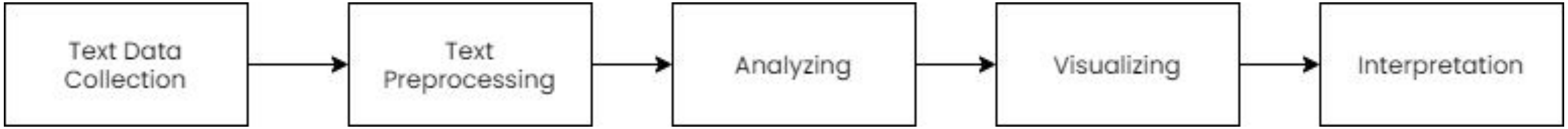
	username	lokasi
1	@turah32	Bandung
2	@juliaaaaa	-
3	@akbarhamdani	Depok
4	@invisble67	Depok
5	@knowuu00	-
6	@undoundo	Tanggerang
7	@maklocantik	-
8	@syafudie	-
9	@danirdn	Purwakarta
10	@locomo23	Bandung

Showing 1 to 10 of 13 entries

Previous 1 2 Next

TELKOM DIGITAL TALENT INCUBATOR 2020

Text Mining Process



Text Data Collection → Text Preprocessing → Analyzing → Visualizing → Interpretation

Organized by:
CAE

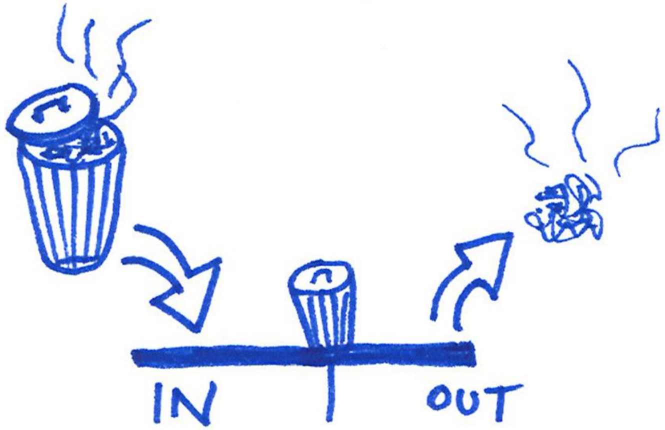
Text Data Collection

Some examples of text data sources:

- Social Media (Twitter, Facebook, WhatsApp, Etc.)
- Website
- News
- Review
- Customer Chat
- Book
- Etc.

Text Preprocessing

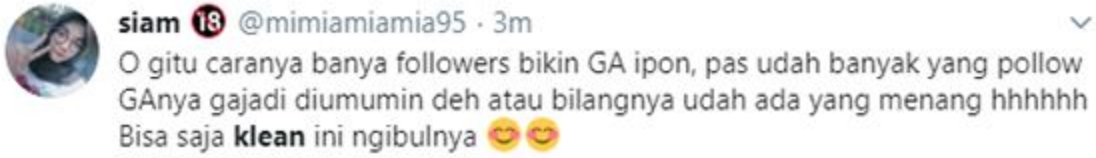
- Text Preprocessing is a process of transforms text into a more digestible form so that machine learning algorithms can perform better.
- One of the Important Task
- Why? **Garbage in Garbage Out**



Common Task in Text Preprocessing

- Text Cleaning
- Feature Extraction

Text Cleaning



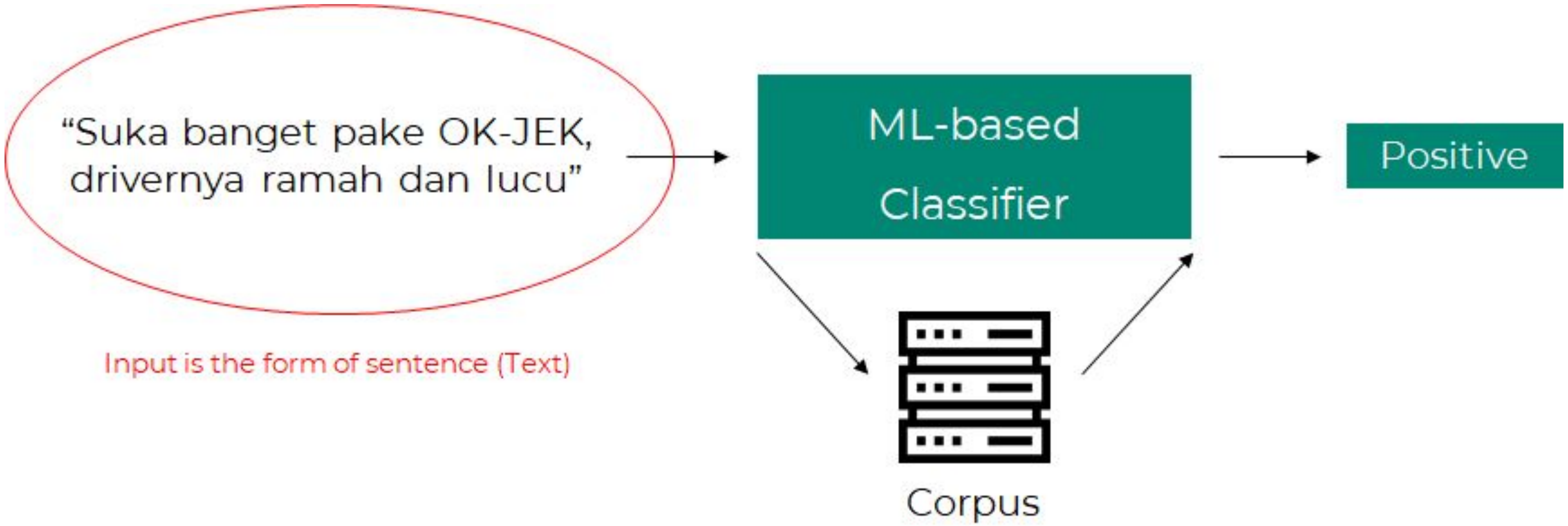
siam 18 @mimiamiamia95 · 3m

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 😊😊

Text Cleaning	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 hhhhhh 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 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.

Feature Extraction

- Machine Learning only process numeric input

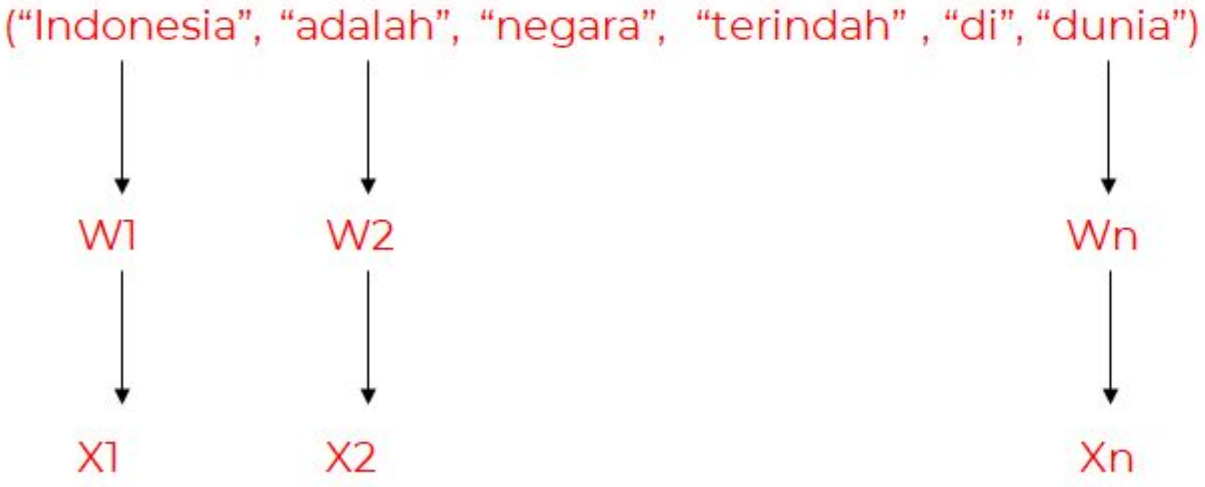


Input is the form of sentence (Text)

Numerical Representation of Text

D = Indonesia adalah negara terindah di dunia.

Tokenize



(“Indonesia”, “adalah”, “negara”, “terindah” , “di”, “dunia”)

W1 W2 Wn

X1 X2 Xn

W = Word (Text)

X = Some numeric encoding of Word

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

TELKOM
DIGITAL TALENT
INCUBATOR 2020

Feature Extraction

Popular Technique:

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

Organized by:
CAE

One Hot Encoding

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

All Word
Direkomendasikan
Banget
Nih
Drivernya
Bau
Ramah
Gak
Bakal
Naik
Lagi

One Hot Encoding

All word	**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

One Hot Encoding

Advantages:

- Simple

Disadvantages:

- Large vocab – Enormous Feature Vector.
- Unordered- Lost All Context.
- Lost Freq. Distribution (Binary).
- Do not capture any semantic information.

Frequency-based Embeddings

Count: Capture how often a word occurs in a document

TF-IDF: Capture how often a word occurs in a document as well as the entire corpus

Co-occurrence: Similar word will occur together and will have similar context

Count

Count, capture how often a word occurs in a document. Ex Count or the Frequency

Advantages:

- Simple

Disadvantages:

- Large vocab – Enormous Feature Vector.
- Lost All Context.
- Do not capture any semantic information.

TF-IDF

TF-IDF: Capture how often a word occurs in a document as well as the entire corpus

Xi = tf(Wi) x idf(wi)

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

Frequently in a single document

Might be important

Frequently in the corpus

Probably a common word

TF-IDF

Advantages:

- Capture Frequency and Relevance

Disadvantages:

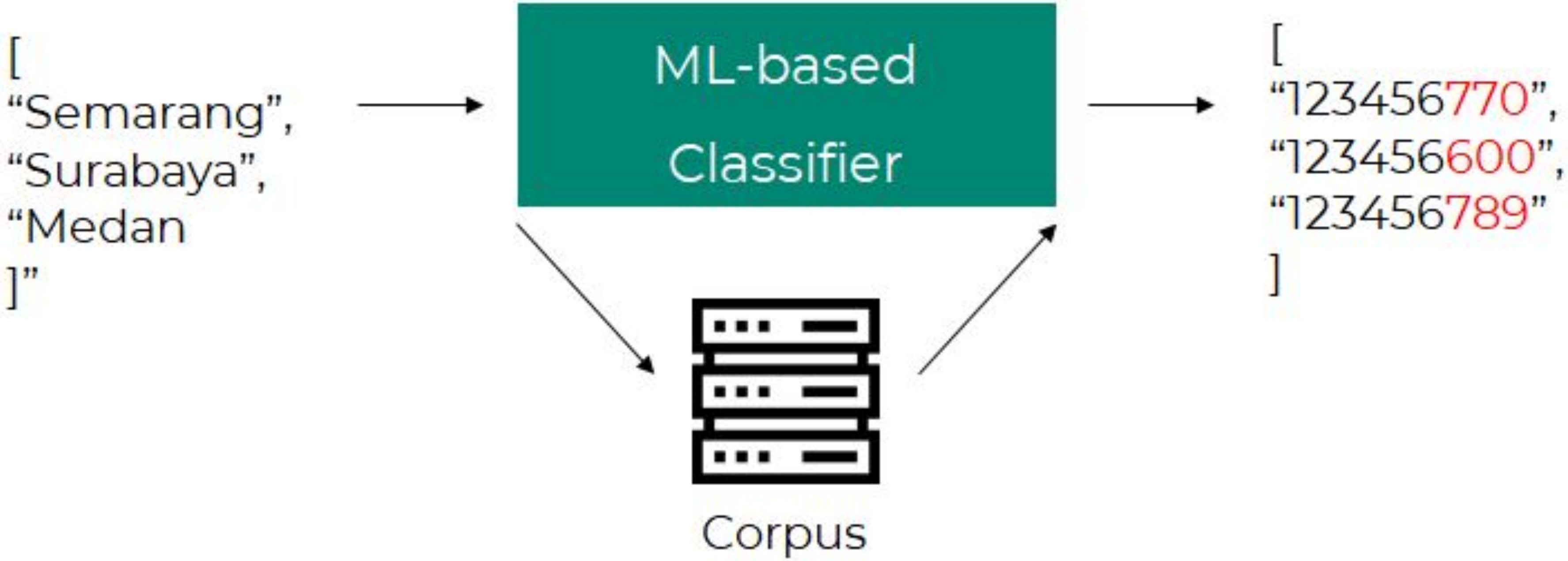
- Context still not captured.

Co-occurrence

- **Co-occurrence:** The number of times two words w1 and w2 have occurred together in a context window
- **Context Window:** A window centered around a word, which includes a certain number of neighboring words.

**Similar word will occur together and will have similar context*

Prediction-Based Embedding



[
“Semarang”,
“Surabaya”,
“Medan
]”

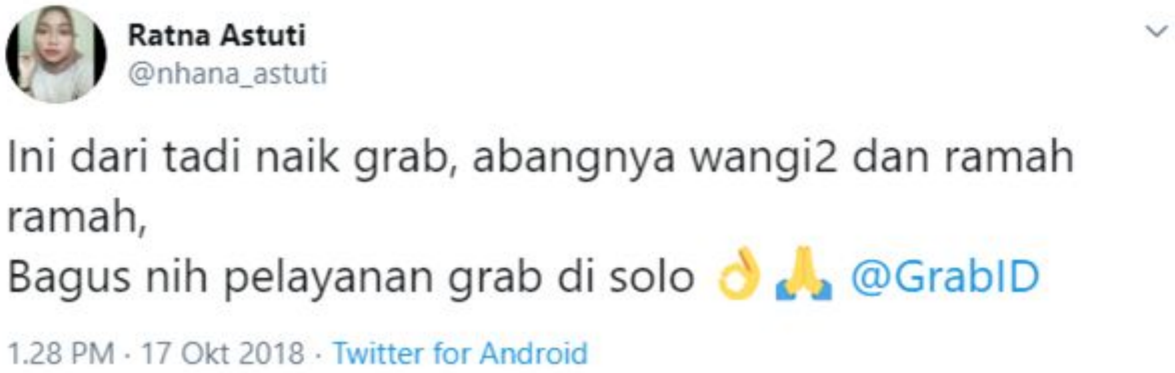
ML-based
Classifier

Corpus

[
“123456770”,
“123456600”,
“123456789”
]

Analyzing: Text Classification

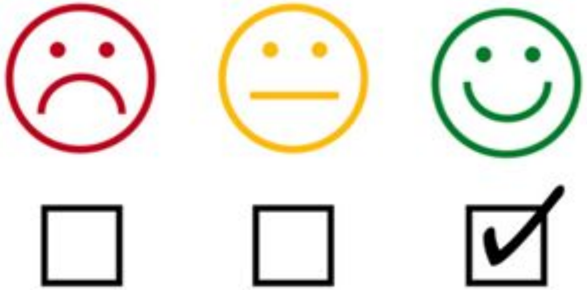
- **Text classification** also known as text tagging or text categorization is the process of categorizing text into organized groups. Some of the most common examples and use cases for automatic text classification is Sentiment Analysis.



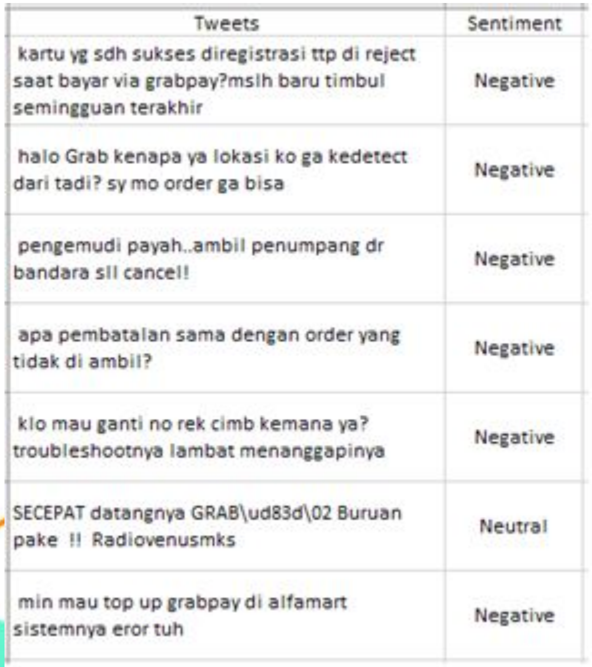
Ratna Astuti
@nhana_astuti

Ini dari tadi naik grab, abangnya wangi2 dan ramah ramah,
Bagus nih pelayanan grab di solo 👌 🙏 @GrabID

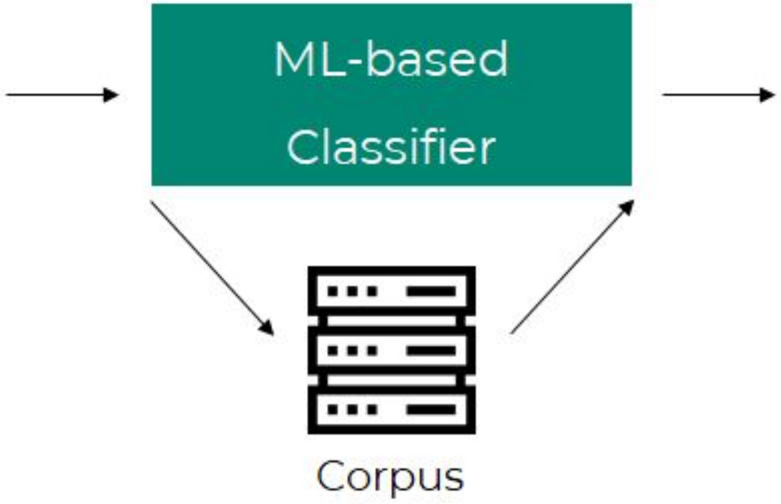
1.28 PM · 17 Okt 2018 · Twitter for Android

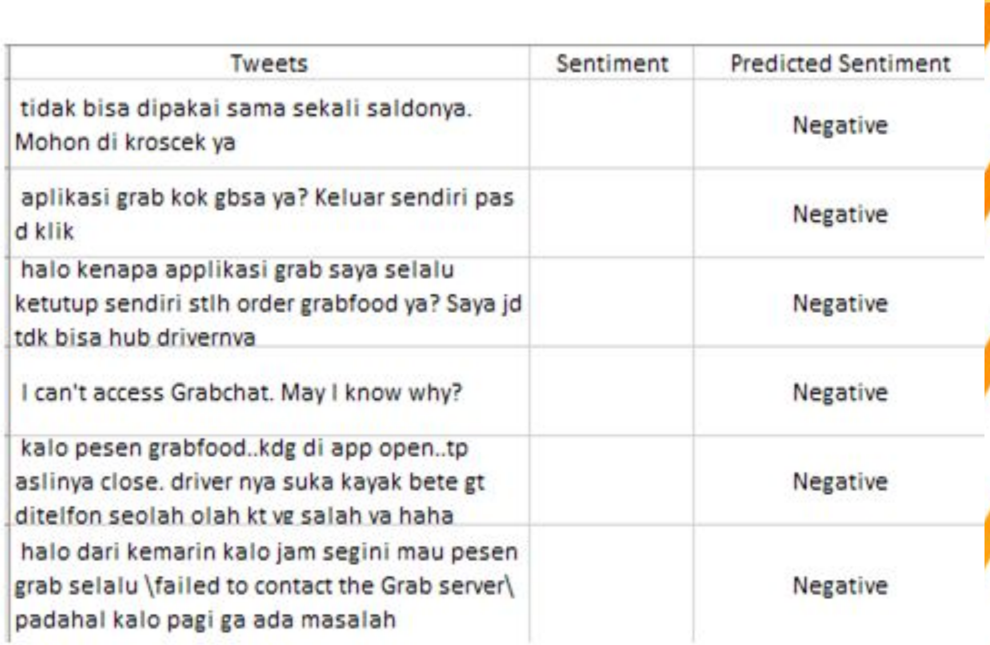


Sentiment Analysis



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



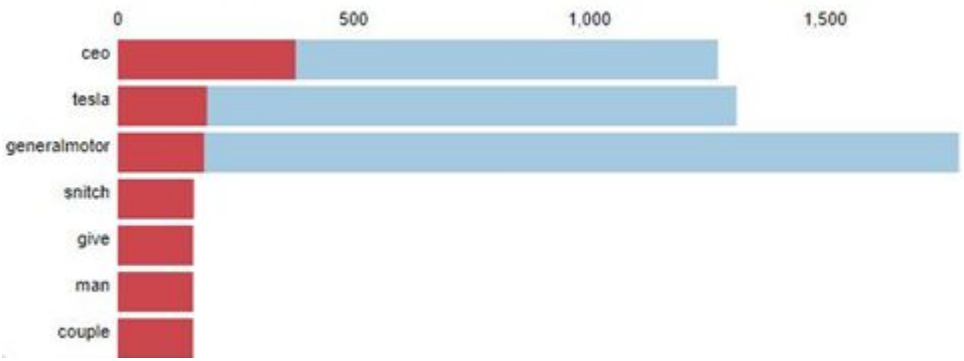


Tweets	Sentiment	Predicted Sentiment
tidak bisa dipakai sama sekali saldonya. Mohon di kroscek ya		Negative
aplikasi grab kok gbsa ya? Keluar sendiri pas d klik		Negative
halo kenapa applikasi grab saya selalu ketutup sendiri stlh order grabfood ya? Saya jd tdk bisa hub drivernya		Negative
I can't access 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 yg 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

TELKOM
DIGITAL TALENT
INCUBATOR 2020

Analyzing: Topic Modelling

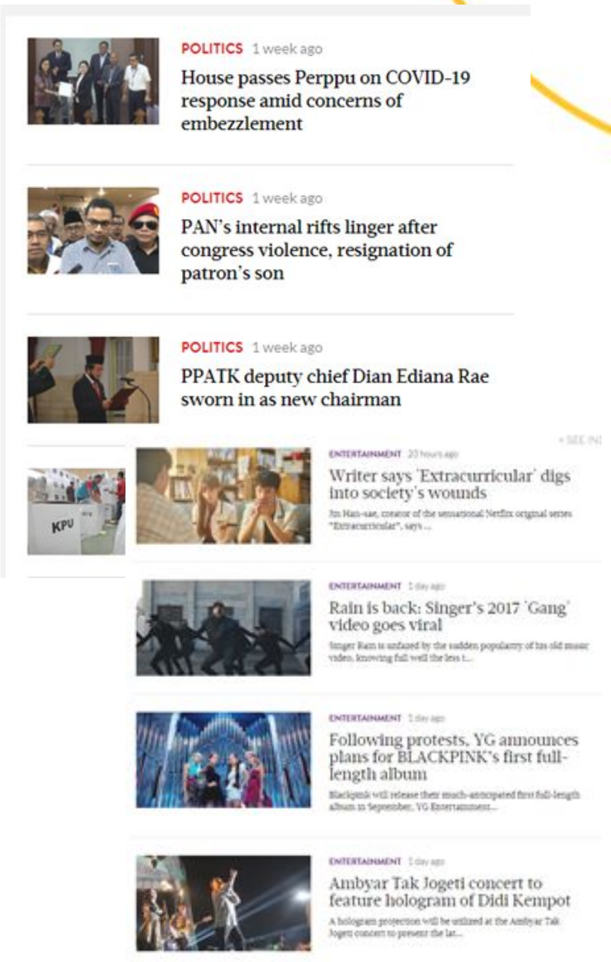
- Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents
- Popular technique:
  - Latent Semantic Analysis (LSA)
  - Probabilistic Latent Semantic Analysis (PLSA)
  - Latent Dirichlet Allocation (LDA)
  - Lda2vec



Organized by:
CAE

Latent Dirichlet Allocation (LDA)

- **Every document is a mixture of topics.**
  - For example, in a two-topic model we could say “Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B.”
- **Every topic is a mixture of words.**
  - For example, we could imagine a two-topic model of Indonesian news, with one topic for “politics” and one for “entertainment.”
  - The most common words in the politics topic might be “President”, “Congress”, and “government”, while the entertainment topic may be made up of words such as “movies”, “television”, and “actor”.
  - Importantly, words can be shared between topics; a word like “budget” might appear in both equally.



POLITICS 1 week ago
House passes Perppu on COVID-19 response amid concerns of embezzlement

POLITICS 1 week ago
PAN’s internal rifts linger after congress violence, resignation of patron’s son

POLITICS 1 week ago
PPATK deputy chief Dian Ediana Rae sworn in as new chairman

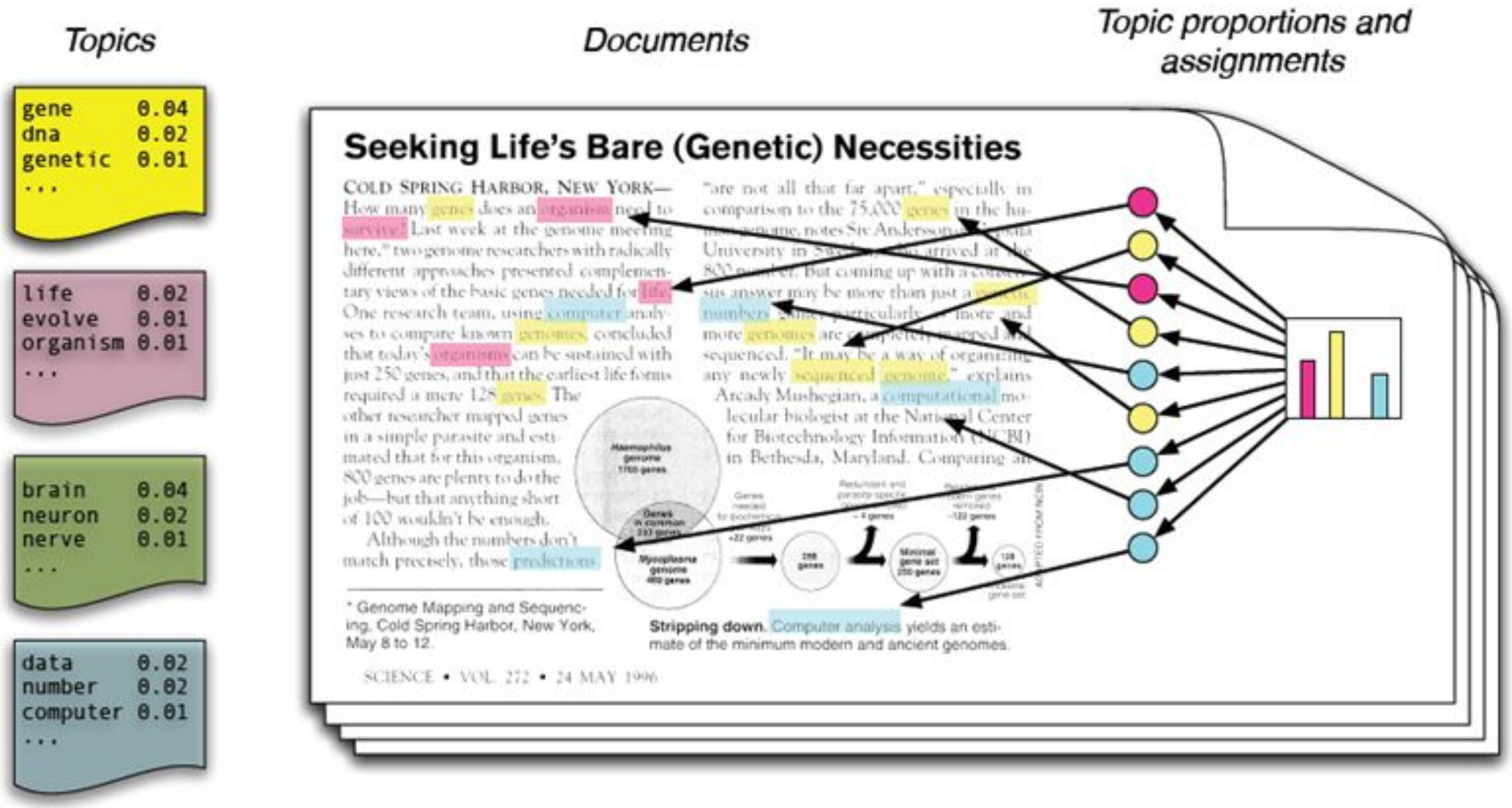
ENTERTAINMENT 20 hours ago
Writer says ‘Extracurricular’ digs into society’s wounds
Jin Han-sae, creator of the sensational Netflix original series “Extracurricular”, says ...

ENTERTAINMENT 1 day ago
Rain is back: Singer’s 2017 ‘Gang’ video goes viral
Singer Rain is unfazed by the sudden popularity of his old music video, knowing full well the less t...

ENTERTAINMENT 1 day ago
Following protests, YG announces plans for BLACKPINK’s first full-length album
Blackpink will release their much-anticipated first full-length album in September, YG Entertainment...

ENTERTAINMENT 1 day ago
Ambyar Tak Jogeti concert to feature hologram of Didi Kempot
A hologram projection will be utilized at the Ambyar Tak Jogeti concert to present the lat...

Latent Dirichlet Allocation (LDA)



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite specific genes removed −4 genes
Minimal gene set 250 genes
Redundant modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Sumber: Probabilistic Topic Models By David M. Blei

Assignment

- Build a model for detecting hate speech using dataset
  https://github.com/ialfina/id-hatespeech-detection
- Use Google Collab or Jupyter Notebook
- Post it on Github