The Challenge of Data Science
Nov 17, 2020

Text Mining and Its Applications in Social Media

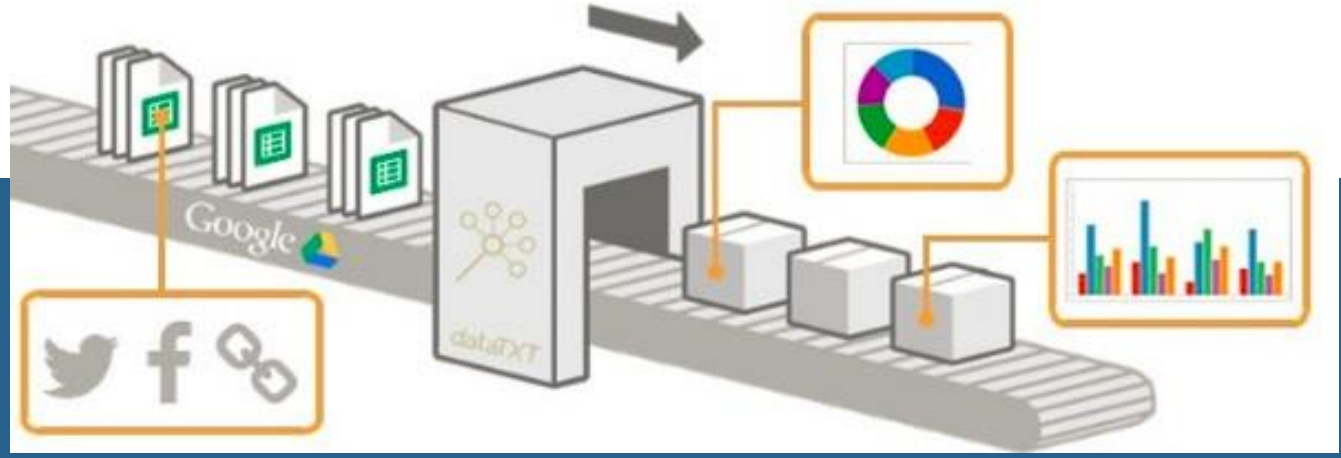
Munawar, PhD

What is text mining ?

- Also known as text data mining
- Process of examining large collections of *unstructured* textual resources in order to generate new information, typically using specialized computer software

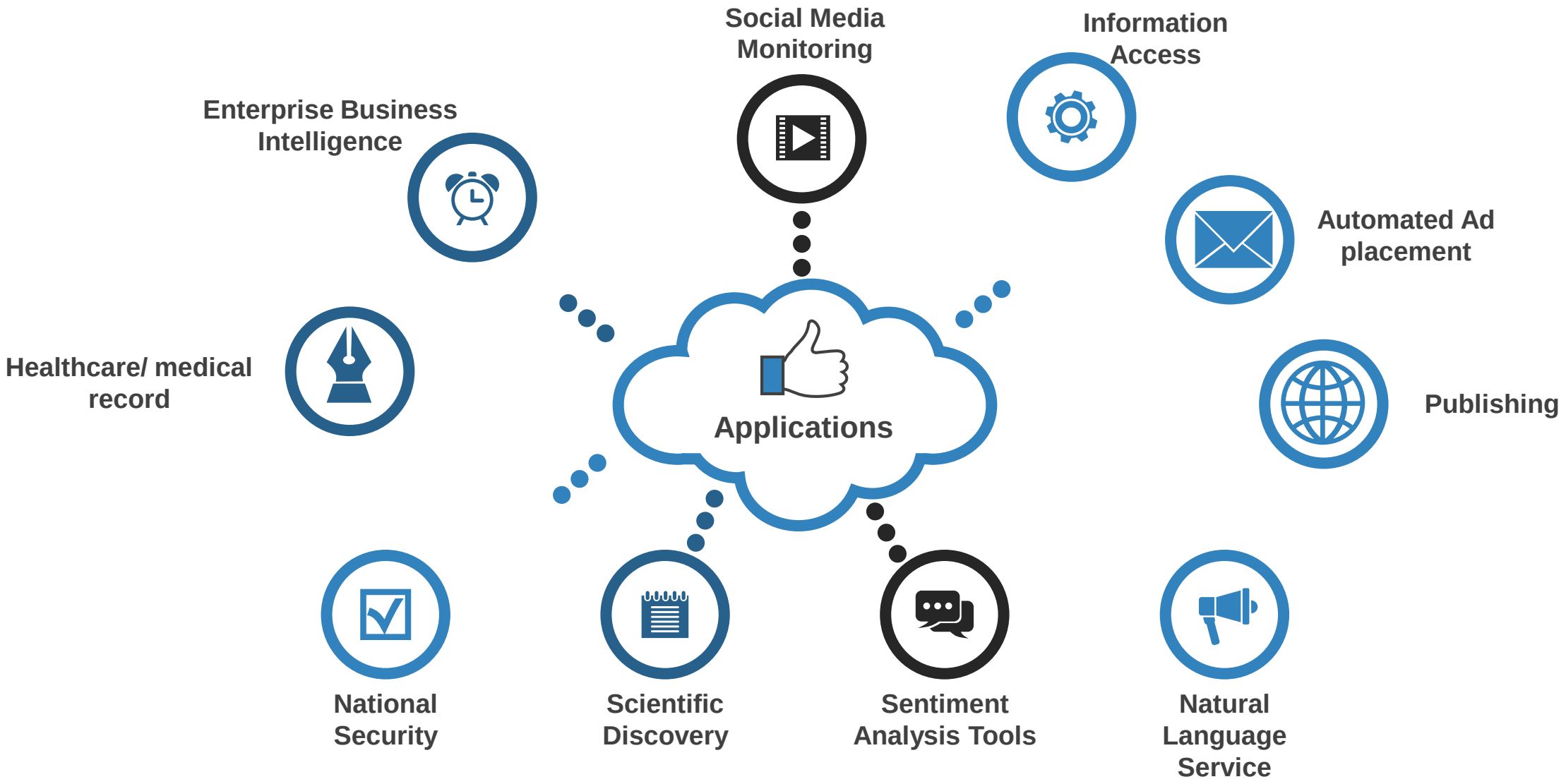
Why do we use text mining ?

- Turn text into data for analysis
- Generate new information
- Populate a database with the information extracted

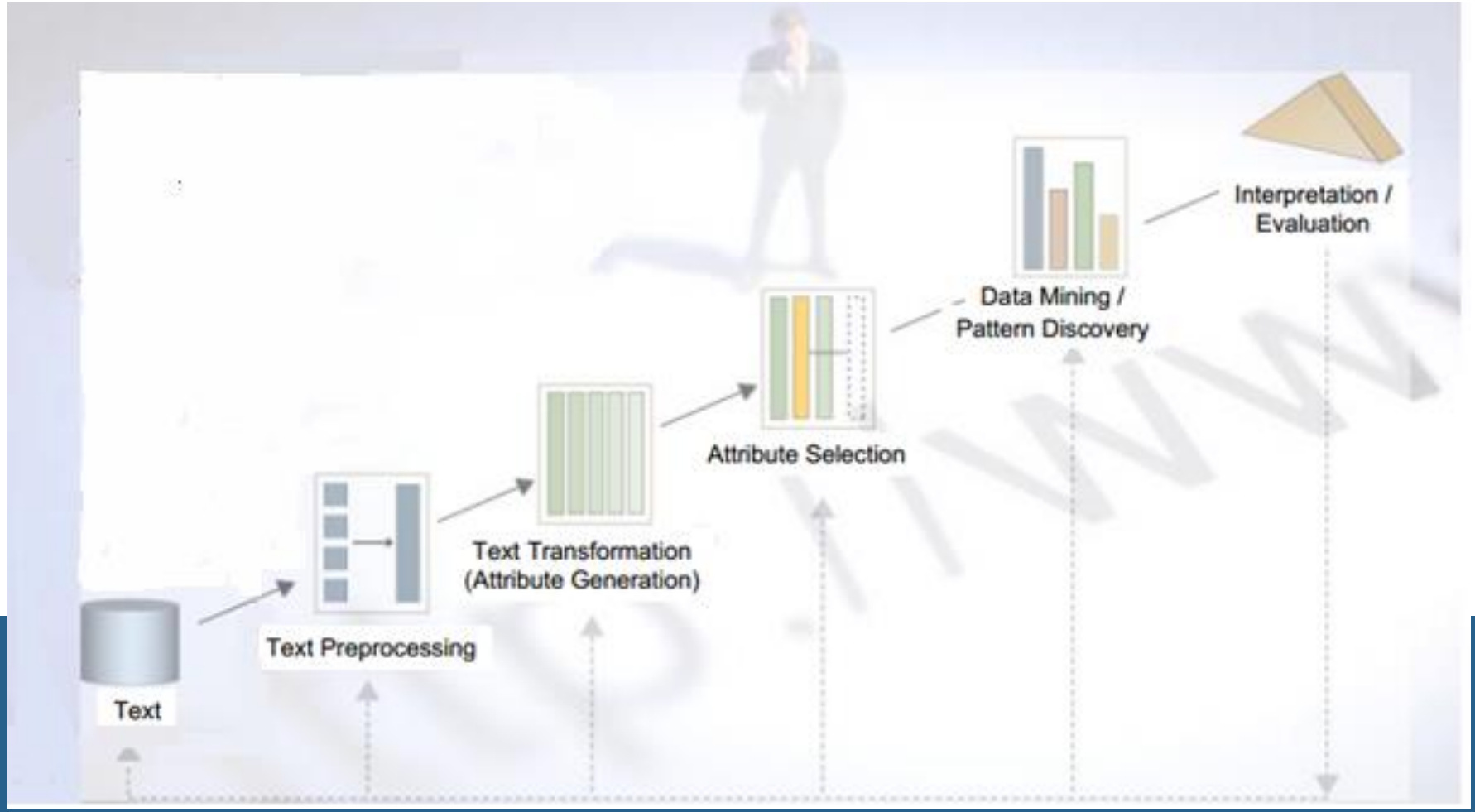


Where Do We See It Being Used?

Applications



Text Mining Process?



Text

- Collect large volume of textual data
- Text Characteristics:
  - High dimensionality w/ tens of thousands of words
  - Noisy data
    - Erroneous data or misleading data
  - Unstructured text
    - Written resources, chat room conversations, or normal speech
  - Ambiguity
    - Word ambiguity or sentence ambiguity

Text Pre-Processing

Text Cleanup

- Normalize texts converted from binary formats (programs, media, images, and most compressed files)
- Deal with tables, figures, and formulas

Tokenization

- Process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

Attribute Generation

- Text document is represented by the words (features) it contains and their occurrences
- Two approaches to generate attributes/document representation:
  - Bag of Words Model, used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature
  - Vector Space Model, used cosine similarity to calculate a number that describes the similarity among documents

Attribute Selection

- Further reduction of high dimensionality
  - Analysts have difficulty addressing tasks with high dimensionality
- Features Selection
  - Select just a subset of the features to represent a document
  - Not all features help
    - Remove stop words
  - Can be viewed as creating an improved document representation

Data Mining

Traditional Data Mining Techniques

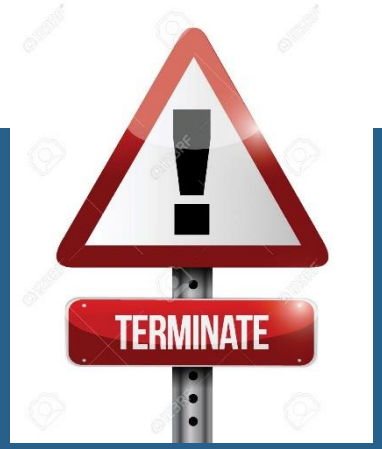
- Classification
- Clustering
- Associations
- Sequential Patterns
- Extract information from the processed text data via data modeling and data visualization (visual maps)

Data Visualization

- Purpose is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables, and charts selected
- makes complex data more accessible, understandable and usable

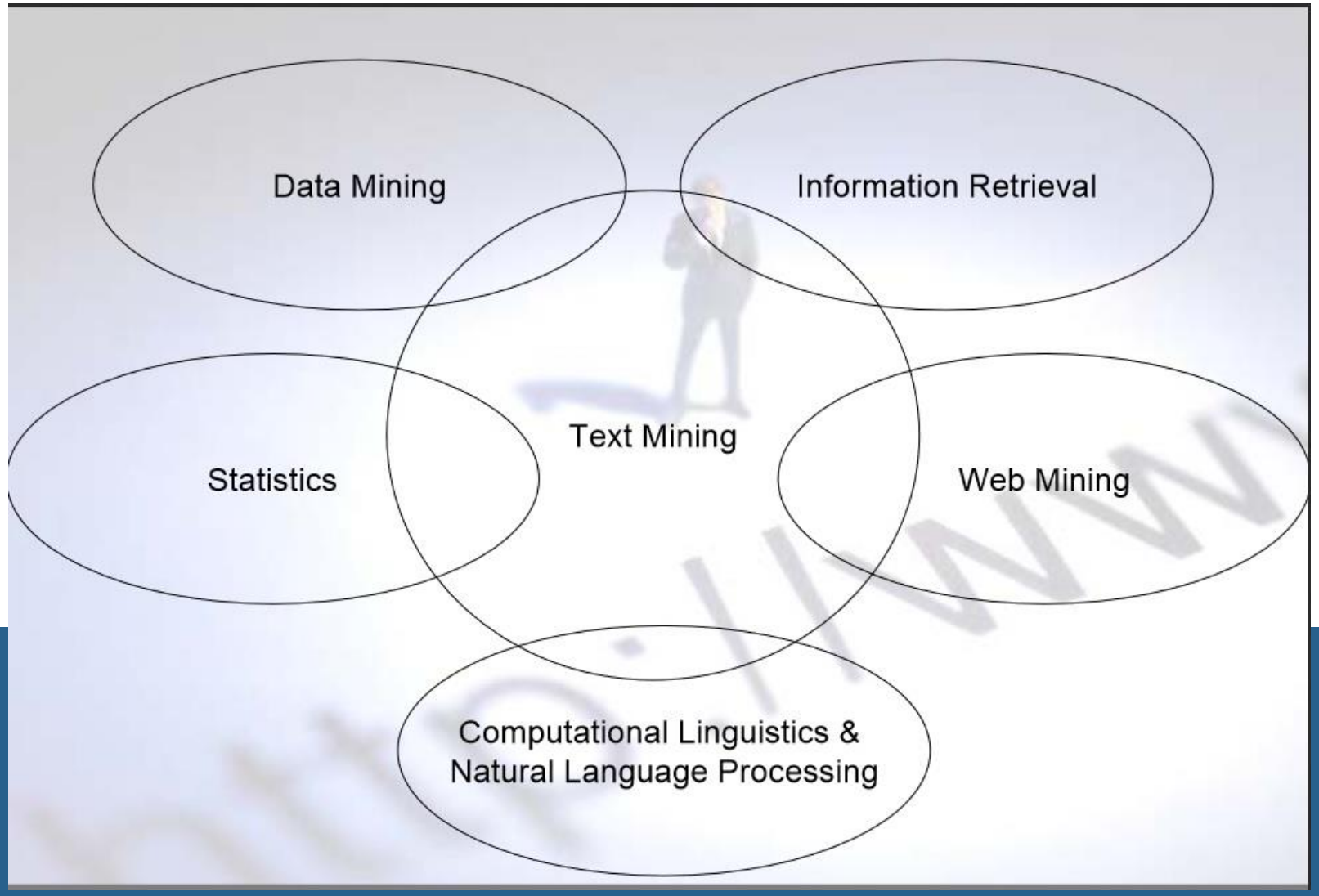
Interpretation/ Evaluation

- Terminate
  - Results satisfied
- Iterate
  - Results not satisfactory but significant
  - the results generated are used as part of the input for one or more earlier stages



Vs.

Text Mining Vs…



Text Mining Vs..

Data Mining:

In Text Mining, patterns are extracted from natural language text rather than databases

Web Mining:

In Text Mining, the inputs are unstructured texts, while web sources' inputs are structured

Text Mining vs Information Retrieval

- New information vs. Web Search
  - No genuinely new information is found
  - The desired information merely coexists with other valid pieces of information
- Hearst's Analogy: "Discovering new knowledge vs. merely finding patterns is like the difference between a detective following clues to find the criminal vs. analysts looking at crime statistics to assess overall trends in car theft"

Text Mining vs …

- Computational Linguistics (CPL) & Natural Language Processing (NLP):
  - CPL computes statistics over large text collections in order to discover useful patterns which are used to inform algorithms for various sub-problems within natural language processing

Text Mining in Social Media

- People use social media to communicate
- Social media provides rich information of human interaction and collective behavior
- Traditional Media vs. Modern Social Media
- Information in most social media sites are stored in text format
- Text Mining can help deal with textual data in social media for research

Distinct Aspects of Text in Social Media

- Textual data provides insights into social networks
- Textual data also presents new challenges:
  - Time Sensitivity
  - Short Length
  - Unstructured Phrases

Aspect #1: Time Sensitivity

- Social media's real-time nature
  Example: some bloggers may update their blog once a week, while others may update several times a day
- Large number of real-time updates from Facebook and Twitter contain abundant information
  Information → detection and monitoring of an event
  Use data to track a user's interest in an event
- A user is connected and influenced by his/her friends
  Example: People will not be interested in a movie after several months, while they may be interested in another movie released several years ago because of the recommendation from his friends

Aspect #2: Short Length

Certain social media websites have restrictions on the length of user's content

- Twitter's 280 characters rule
- Windows Live Messenger's 128 character personal status

Short Messages → people become more efficient with their participation in social media applications

Short Messages also bring new challenges to text mining

Aspect #3: Unstructured Phrases

- Variance in quality of content makes the tasks of filtering and ranking more complex
- Computer software have difficulties to accurately identify semantic meaning of new abbreviations or acronyms

Applying Text Mining in Social Media

Certain aspects of textual data in social media presents great challenges to apply text mining techniques

Event Detection

Event Detection aims to monitor a data source and detect the occurrence of an event that is captured within that source

- Monitor Real-Time Events via Social Media
- Example: Detecting earthquake when people are posting live-situation through microblogging like Twitter & Facebook

Improve traditional news detection

- Large number of news are generated from various new channels, but only few receive attention from users
- Researchers proposed to utilize blogosphere to facilitate news detection

Collaborative Question Answering

- Collaborative question answering services bring together a network of self-declared “experts” to answer questions posted by other people
- Through text mining, a tremendous amount of historical QA pairs have built up their databases, and this transformation gives users an alternative place to look for information, as opposed to a web search
- The corresponding best solutions could be explicitly extracted and returned

Social Tagging

- A method for Internet users to organize, store, manage and search for tags / bookmarks (also as known as social bookmarking) of resources online
- Social Tagging vs. File Sharing
- Through text mining, it helps to improve the quality of tag recommendation
  - Facebook’s tag recommendation of a photo
- Utilize social tagging resources to facilitate other applications
  - Web object classification, document recommendation, web search quality

Concerns For Text Mining

- Text in unstructured documents is hard to process
- The information one needs is often not recorded in textual form
- We do not have programs that can fully interpret text. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do

Future Of Text Mining

- As most information (common estimates say over 80%) is currently stored as text
- This includes emails, newspaper or web articles, internal reports, transcripts of phone calls, research papers, blog entries, and patent applications
- Thanks to the web and social media, More than 7 million web pages of text are being added to our collective repository, daily
- We can now begin to see the usefulness of software that can process between 15,000- 250,000 pages an hour, compared to a mere 60 pages for humans
- Text mining is believed to have a high commercial potential value

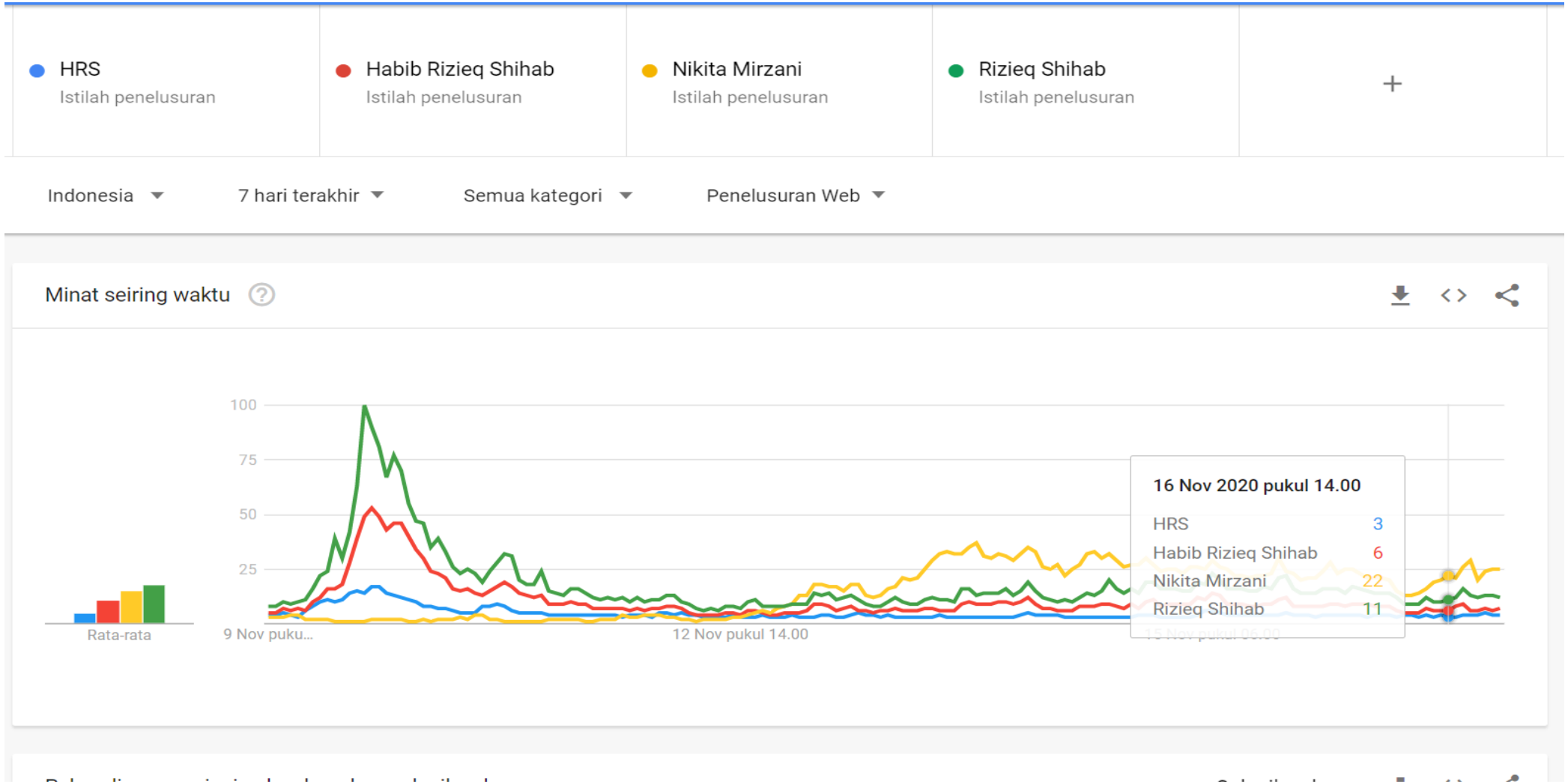
Thank You

moenawar@gmail.com

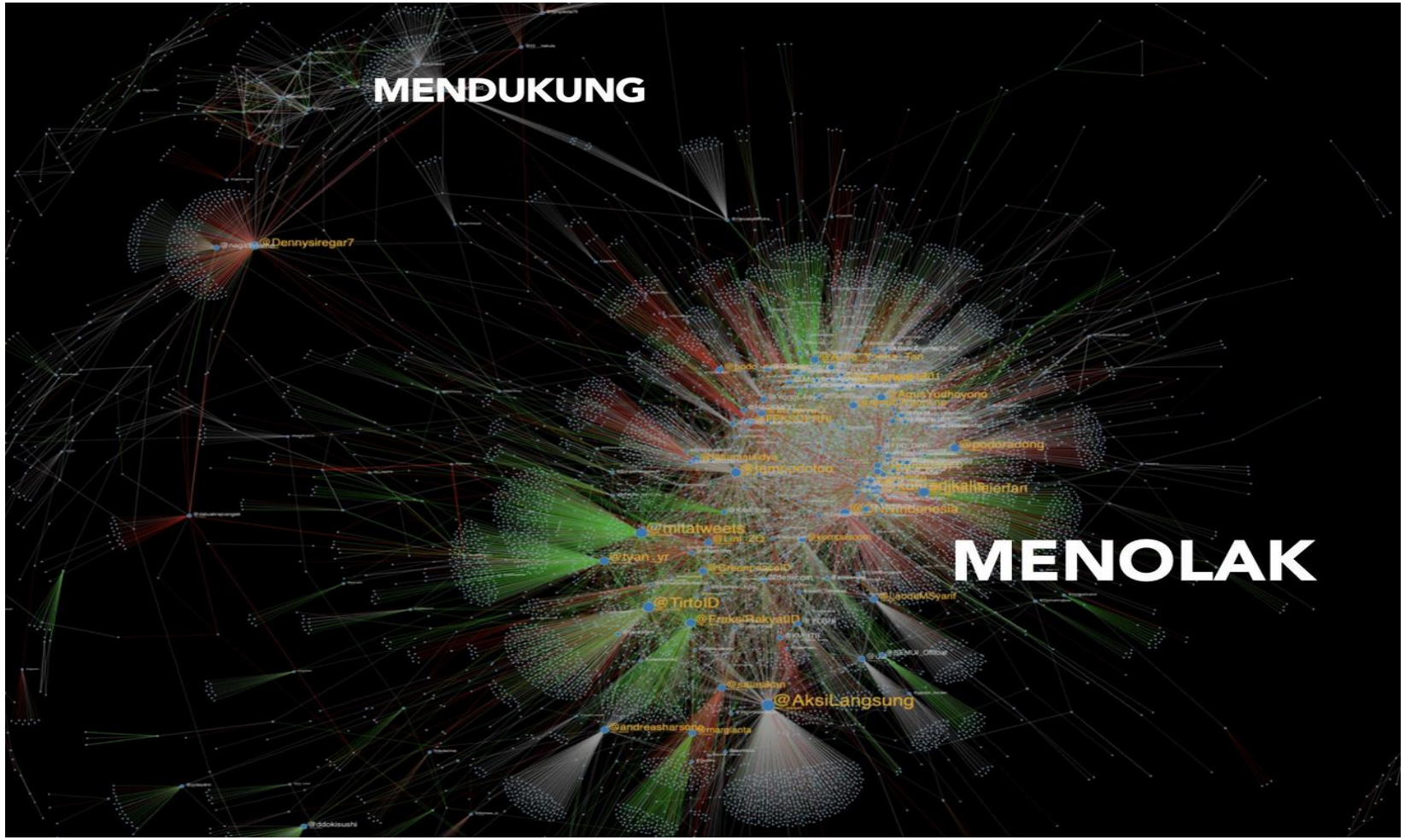
Free Trial …

https://www.vicinitas.io/free-tools/download-user-tweets

Contoh Penerapan Text Mining



Peta SNA Omnibus Law 28 Sep – 5 Okt



Peta SNA Omnibus Law 23 Okt – 26 Okt