Text Mining and Sentiment Analysis

Prof. Annamaria Bianchi
A.Y. 2024/2025

Lecture 1

17 February 2025

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Scienze Economiche

Outline

Structured vs unstructured data

Primary vs secondary data

Data sources

Main text sources

Structured vs unstructured data

- The term **structured data** refers to data that is formatted and organised in a data structure so that elements can be addressed and accessed in various ways to make better use of the information (e.g. spreadsheet and relational databases - database that organizes data into rows and columns, which collectively form a table where the data points are related to each other). Variables are clearly defined. It can consist of numbers and text.
- Typical examples of structured data are names, addresses, credit card numbers, geolocation, and so on.

- **Unstructured data** is basically the opposite of structured: unstructured data is not organised in a format that makes it easy to access and process. Even though unstructured data may have a native, internal structure, it's not structured in a predefined way. Data is stored in its native format. The choice of variables is not clear.
- Typical examples of unstructured data are rich media, (unprocessed) text, social media activity, surveillance imagery, and so on.
- The amount of unstructured data is much larger than that of structured data. Unstructured data makes up 80% or more of all enterprise data, and the percentage keeps growing.

Structured vs unstructured data

- **Semistructured data** is a third category that falls somewhere between the two extremes. While it has some organization, it doesn't have enough structure to meet the requirements of a relational database.
- It's a type of structured data that does not fit into the formal structure of a relational database. But while not matching the description of structured data entirely, it still employs tagging systems (metadata tagging) or other markers, separating different elements and enabling search.
- A typical example of semistructured data is smartphone photos. Every photo taken with a smartphone contains unstructured image content as well as the tagged time, location, and other identifiable (and structured) information. Semi-structured data formats include JSON, HTML, and XML file types.

Structured vs Unstructured data: key differences

1) **Format**

Structured data has a strict, predefined data model. Unstructured data does not have a predefined format.

2) Storage

Structured data storage systems have rigid schemas, such as those in relational databases or data warehouses. Unstructured data is often stored in its native format in nonrelational databases or data lakes.

3) Use cases

Organizations can use both structured and unstructured data across artificial intelligence (AI) and analytics use cases. Structured data is often used in machine learning (ML) and drives ML algorithms . Unstructured data is often used in natural language processing (NLP) and is a rich and diverse data source for generative AI (gen AI) models.

Structured vs Unstructured data: key differences

4) **Complexity**

Structured data is easier to manipulate and analyze for general business users with traditional tools. Unstructured data can be more complex and requires specialized skills and tools to parse and analyze.

Primary vs Secondary data

As the name suggests, **primary data** are first-hand information collected by the researcher. The data so collected are pure and original and collected for a specific purpose.

Methods of collecting primary data:

- Surveys
- Censuses
- Observation methods
- Experimental methods
- Interview methods

Primary vs Secondary data

Secondary data is the data which has already been collected and then reused again for some valid purpose.

Internal sources: These types of data can easily be found within the organization such as market record, sales record, transactions, customer data, accounting resources, etc.

External sources: administrative sources and sources nowadays identified as ‘Big data’

Mindful of the costs and response burden involved in the collection of primary data, more and more organizations aim to maximize the use of secondary data for statistics purposes.

To which type of data sources does Text Mining fit?

Text mining transforms unstructured data into structured data

Text mining can be useful each time textual data naturally fits as an input

It applies mainly to **secondary data**, but one can also find applications in the context of **primary data** sources

Main Text Sources

Web scraping

- **Web scraping**, **web harvesting**, or **web data extraction** is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol (HTTP) or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet.
- The R programming language (using proper package) enables you to simply and efficiently scrape web pages.

Main Text Sources

Application Program Interfaces (APIs)

- An **application programming interface** (**API**) is a computing interface that defines interactions between multiple software intermediaries. It defines the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conventions to follow, etc. It can also provide extension mechanisms so that users can extend existing functionality in various ways and to varying degrees. An API can be entirely custom, specific to a component, or designed based on an industry-standard to ensure interoperability.
- APIs are the links between applications allowing information to be shared.
- R packages devoted to specific APIs.

File sources

- Files in different formats: .doc, .txt, .pdf, .csv, and .xml.
- Different R functions used for different formats