IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS OF UNSTRUCTURED DATA USING ARTIFICIAL INTELLIGENCE

Author:

Ravi Rajnikant Oza

Assistant Professor: Bhavans Shree H. J. Doshi Info. Tech. Institute, Jamnagar, Gujarat

PhD Scholar: Department of Computer Science, Saurashtra University, Rajkot, Gujarat

Corresponding Author:

Dr. Dipti H. Domadiya

Associate Professor

National Computer College, Jamnagar

Abstract

Wealth of information that is stored in unstructured data, but unstructured data lacks simple recognizable structure such that it cannot be easily used by computer software. It does not have pre-defined model or structure, therefore it is not stored in relational database. Unstructured data could be machine generated or human generated; it is stored in the native format along with the associated metadata. To analyze unstructured data we must use cotemporary technology such as artificial intelligence and machine learning. While analyzing unstructured data we must consider metadata, Natural Language Processing (NLP), Image analysis and Data Visualization.

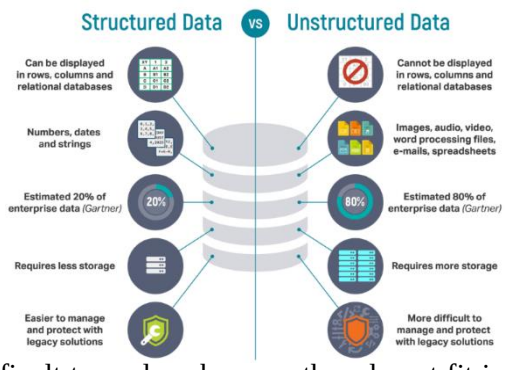
Keywords

Unstructured Data, Artificial Intelligence, Machine Learning, Natural Language Processing, Image Analysis, Data Visualization

1. Introduction

A vast majority of the data now a day in digital world is of unstructured type. Unstructured data, also known as unstructured information, refers to information that lacks a predefined data model or is not organized in a specific way. Unstructured data, also known as unstructured information, refers to information that lacks a predefined data model or is not organized in a specific way.

Unstructured data is the data, which cannot be stored in a relational databases (RDBMS) traditional structure. It is sometimes referred to as qualitative (estimated) data. We can develop meaning from it, but it may also be very unclear and difficult to analyze.



Unstructured data does not follow to a specific data model, it does not have simple recognizable structure such that it cannot be used by computer software effortlessly. It does not have pre-defined model or structure, therefore it does not proper in relational database.

The file type is known when working with unstructured data, but the file content is unknown. Such files' contents are self-explanatory. They are difficult to analyze because they do not fit in a relational database.

According to IDC (International Data Corporation), 80% of the world's data will be unstructured by 2025. Organizations have recently begun to pay attention to unstructured data, in particular the wealth of information that is stored in unstructured data and that is concealed behind unstructured data, and they now want to take advantage of it. [1]

In order to derive insightful information from unstructured data and make pertinent business decisions, organizations must do this analysis. The decisions made based on these results will influence how customers feel about future planning, help to understand customer needs, and help to identify contributions that will better meet those needs.

Machine learning algorithms and natural language processing techniques are used to analyze unstructured data to train a model to work for a specific need.

2. Sources of unstructured data [2]

Broadly there are two main sources of unstructured data, the machine generated and human generated.

1. Machine generated unstructured data:
    a. Satellite images (google maps to detect live traffic)
    b. Data from medical imaging devices (ECG, MRI, etc.,)
    c. CCTV surveillance data
    d. Atmospheric data (temperature, humidity level, speed of wind etc.,)
    e. Radar data (live tracking of flights)
2. Human generated unstructured data
    a. E-mails (text, images, gif etc.,)
    b. Text messages (text, images, gif, multimedia etc.,)
    c. Social media data (tweets and post with text and multimedia)
    d. Audio recording
    e. Office documents (word, excel, pdf etc.,)
    f. Images (photos taken from smartphone and digital camera)

3. Analysis of unstructured data

We must first determine which source data is necessary for the data analytics before we can move forward[3]. Unstructured data can be found in a variety of formats, including webpages, tweets, posts, videos, audio, and images; we must only analyze the data that is pertinent to our particular need. While there are various sources of data accessible, it is vital to choose the ones that are most relevant and applicable to our particular requirements.

It may be necessary to remove unnecessary data or anything that is not fit for purpose. It is also necessary to choose appropriate technology and tools for gathering, refining, storing, and managing data. Unstructured data is saved in its original format, along with the corresponding metadata.

Once all of these elements are established, we can devise a methodology for processing and analyzing the unstructured data. Now, let us examine the potential steps involved in analyzing unstructured data.

1. Metadata

Data about data is known as metadata, means the data itself provide information related to its data. It is very vital to store, manage and analyze the unstructured data. Object storage systems enable the storage of vast quantities of disorganized data, which is typically written just once and then accessed multiple times [4]. What is Object storage exactly? In brief, it is a storage for unstructured data that removes the scaling limitation of traditional file storage. Unlimited scale is the motive that object storage is the storage of cloud.

The purpose of object storage is to store things like Facebook photos and videos, Gaana music files (songs), or files in online collaboration services like outlook and google drive.

The storage mechanism involves the creation of objects that consist of data, metadata, and a unique identifier. There are various metadata fields that can be filled in for most file types.

Each image we take with a digital camera or smartphone has metadata like the date, time, filename, and location attached to it.

Metadata such as the blog title, writer of the blog, URL of blog, blog issuing date, labels (tags), and category are included in every blog post. Metadata on a page incorporates things like the page title, URL of the page, page portrayal, and symbol.

We can create additional custom metadata fields based on your needs to describe the nature or contents of the unstructured data in addition to these standard fields. In this way, metadata can make further search and analysis easier.

Each enterprise must establish its own metadata standards because there are presently no industry-wide ones. Effective metadata use facilitates data organization, automation, policy enforcement, and data visibility. Even though it's ideal to attach metadata at the time the data is created, sometimes it needs to be done after the fact.

2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of computer science, and more specifically, artificial intelligence (AI), that focuses on enabling computers to comprehend written and spoken language in a manner similar to humans. [5]

Natural Language Processing (NLP) is a machine learning technique that helps to analyze the meaning of unstructured textual data. NLP attempts to replicate the human mind's ability to process natural languages like Hindi, Gujarati, English, and others.

Even when documents do not adhere to a standard template, NLP can determine the meaning of text data in a context.

Let's examine some NLP models for handling unstructured text.

- ✓ Bag of Words : This technique simply involves counting the frequency of a word or phrase in the text, without considering its semantic context.
- ✓ Tokenization : It is a process used to divide strings into words and sentences, while disregarding punctuation, commonly used in natural language processing.
- ✓ Stop words removal : It is a technique that involves removing articles and prepositions from the text that do not add any significant value to the NLP process. These words include "a," "the," "and," and "to." A basic list of stop words is usually used to start the analysis, which can be refined depending on the specific goal of the analysis.
- ✓ Stemming : It is a natural language processing technique that involves adding a prefix or suffix to a word in order to group different variations of the same word together. For instance, "sitar" and "sitarist" can be organized using stemming.
- ✓ Lemmatization : It is a process that involves converting words to their dictionary form, also known as their "lemma." This involves removing tenses, so words like "teaching" and "taught" both become "teach." Synonyms are also combined during lemmatization; for example, "home" becomes "house." Lemmatization considers the context of the word, as the same word may have multiple meanings depending on the context in which it is used.
- ✓ Topic modeling: A text-mining tool called topic modelling can help to identify themes in the text and locate word clusters associated with several themes. The words that appear frequently in a document can be used to identify its themes and meaning. In order to determine the topic of unstructured data, topic modelling involves adding up words and combining related patterns. A topic model group can determine whether a pattern is similar by looking at factors like word frequency and word distance, as well as the words and expressions that appear the most frequently.

3. Image Analysis

Let's focus on images now since unstructured data not only includes text but also a large portion of image content. We need to realize that images contain a wealth of information. If we have an image of a student learning a subject in a classroom, for instance, we can use image analysis to determine the student's behaviour throughout the lesson in order to identify human behaviour by analyzing their gestures.

New systems have been created that can retrieve images by analyzing unstructured data, such as MRI images that correspond to a specific brain volume or X-rays of the backbone that match a particular backbone image. These systems utilize feature extraction and similarity matching techniques to identify similar images based on an input image [6].

It is feasible to have automatic or intelligent vehicles on the road thanks to AI-based image analysis because they can identify objects on the road and determine their locations.

4. Data Visualization

Data visualization is a technique for putting data in a graphical format so that it is much easier to understand. Data that has been graphically visualized is easier to understand, encourages interaction, and reveals a wealth of new information. Data visualization makes it possible for more people to understand complex data structures that would otherwise be impossible to understand.

For example, I am having a Samsung Smartphone, a feature called "Digital Wellbeing" displays an app dashboard where users can see how long each application was open. By swiping to different screens, users can also view breakdowns by day, by hour, and by app, all of these is displayed in a well design data visualization manner. I can easily understand how much time spent with a certain app through the week.

The most appropriate visualization techniques for the application and the intended audience can be used to present unstructured data that has been processed using methods like topic modelling, sentiment analysis, information extraction from images and data visualization.

4. Challenges with unstructured data

Since the diagram makes it clear that 80% of data is unstructured, it is essential to understand what that data means, it stores wealth of information in unstructured data and it is hidden. This can only be done by analyzing the data, which is aided by artificial intelligence and machine learning.

But, unstructured data presents a variety of challenges, from collection to storage to decision-making. Such data cannot be analyzed easily with the present database as most of the current data analytics databases are intended for structured data. Consequently, data analytics engineers have to look for the newer methods to locate, extract, organize and store unstructured data.

Here is a list of the difficulties encountered when working with unstructured data. [7]

- ✓ Conventional system such as RDBMS is not capable to analyze this data.
- ✓ Since unstructured data grows continuously, it has a very high volume of data, therefore the larger the dataset, the harder it is to store and manage, to deal with such data, we need systems that can process large volume of data efficiently.
- ✓ When we analyze the unstructured data, the relevance is one of the biggest issues since the data analytics models cannot make distinction between action and relationship.
- ✓ Unstructured data is frequently not available in high quality, particularly when it comes to images and videos.

5. Conclusion

We should make use of contemporary techniques like artificial intelligence and machine learning to analyses unstructured data in order to extract the wealth of information hidden within it because the nature of unstructured data makes it difficult to analyses and conventional database analytics would not be able to do it. Data analytics engineers are currently creating different algorithms and testing them on various datasets to gain insightful information. My own research work is related to this, where I proposed to develop a model that identifies gestures to determine human behaviour from unstructured data.

6. References

[1] "Structured vs Unstructured Data | Semi Structured Data." https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/ (accessed Dec. 05, 2022).

[2] *What is Unstructured Data | How it is stored | Sources | Real Life Examples | Amit Thinks | 2022*, (Apr. 10, 2022). Accessed: Dec. 05, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=PDeZRbcrzWA

[3] "7 Steps to Analyze Unstructured Data," *Orchestrate Blog*, Feb. 16, 2015. https://www.orchestrate.com/blog/7-steps-to-analyze-unstructured-data/ (accessed Dec. 05, 2022).

[4] "Object storage," *Wikipedia*. Nov. 29, 2022. Accessed: Dec. 02, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Object_storage&oldid=1124607101

[5] "What is Natural Language Processing?," Aug. 16, 2021. https://www.ibm.com/cloud/learn/natural-language-processing (accessed Dec. 02, 2022).

[6] T. T. Group, "How to Analyze and Process Unstructured Data," *Treehouse Tech Group*, Dec. 04, 2019. https://treehousetechgroup.com/how-to-analyze-and-process-unstructured-data/ (accessed Dec. 04, 2022).

[7] "Big Data and the Challenge of Unstructured Data." https://www.ciklum.com/blog/big-data-and-the-challenge-of-unstructured-data (accessed Dec. 05, 2022).